VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

GENOTYPIZACE U MAKAKŮ VE VÝZKUMU INFEKCE VIREM HIV

MHC AND KIR GENOTYPING OF MACAQUES IN HIV INFECTION RESEARCH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — Jan Matula
AUTHOR

VEDOUCÍ PRÁCE — Mgr. Ing. Karel Sedlář
SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Student: Jan Matula ***ID:*** 164212
Ročník: 3 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Genotypizace u makaků ve výzkumu infekce virem HIV

POKYNY PRO VYPRACOVÁNÍ:

1) Zpracujte literární rešerši nejpoužívanějších sekvenačních technologií se zaměřením na pyrosekvenování. Zaměřte se na metodu sekvenování amplikonů a princip multiplexace vzorků. 2) Prostudujte možnosti využití sekvenačních dat pro možnost genotypizace při výzkumu infekčních nemocí. 3) Ve vhodně zvoleném jazyce vytvořte funkce pro demultiplexaci vzorků a základní kvantitativní analýzu surových sekvenačních dat. Data pro ověření metody poskytne Centrum pokročilých studií, Univerzita obrany. 4) Balíček doplňte o funkce umožňující identifikaci sekvencí pomocí vhodně zvolené referenční databáze. 5) Upravte stávající nástroje pro statistické hodnocení, případně navrhněte nové tak, aby bylo možné dále analyzovat předzpracovaná data. 6) Proveďte statistické zhodnocení pro jednotlivé vzorky analyzovaných dat a výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] WISEMAN, R. W., et al. Major histocompatibility complex genotyping with massively parallel pyrosequencing. Nature Medicine. 2009-10-11, vol. 15, issue 11, s. 1322-132.

[2] BABIK, W., et al. New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHCsystem. Molecular Ecology Resources. 2009, vol. 9, issue 3, s. 713-719.

Termín zadání: 6.2.2017 ***Termín odevzdání:*** 2.6.2017

Vedoucí práce: Mgr. Ing. Karel Sedlář
Konzultant:

prof. Ing. Ivo Provazník, Ph.D.

předseda oborové rady

Abstrakt

Moderní výzkum virových onemocnění je závislý na analýze genetických dat a to nejen samotných virových sekvencí, ale i specifických receptorů, které na virové onemocnění reagují. Tato práce představuje balíček nástrojů vytvořený v programovacím jazyce R s využitím nadstavby Bioconductor pro zpracování dat produkovaných next generation sekvenátory. Balíček využívá pokročilý algoritmus SSAHA pro porovnávání neznámých sekvencí DNA s referenční databází. Funkčnost balíčku je demonstrována na datech získaných sekvenováním MHC a KIR receptorů HIV pozitivních makaků na platformě Roche 454.

Klíčová slova

HIV, genotypizace, MHC, SSAHA, R, databáze

Abstract

Modern research of viral diseases relies on genomic data processing. Not only is the sequence of a virus important, genomic sequence of specific receptors in affected organisms also plays an important role. In this paper, a novel package for processing of next generation sequencing data in infectious disease written using R/Bioconductor language is proposed. Functionality of the package, including implementation of advanced SSAHA algorithm for fast database searches, is demonstrated using genotyping of genes for MHC and KIR receptors of HIV positive macaques.

Key words

HIV, genotyping, MHC, SSAHA, R, databases

MATULA, J. *Genotypizace u makaků ve výzkumu infekce virem HIV*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2017. 46 s. Vedoucí bakalářské práce Mgr. Ing. Karel Sedlář.

Prohlášení

Prohlašuji, že svou bakalářskou práci na téma Genotypizace u makaků ve výzkumu infekce virem HIV jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

Poděkování

Děkuji svému vedoucímu Mgr. Ing. Karlu Sedlářovi za odbornou pomoc a cenné rady při vypracovávání této bakalářské práce. Další díky patří panu prof. MUDr. Pavlu Boštíkovi, Ph.D. za poskytnutí dat k analýze.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Pokrok v rychlosti a kvalitě čtení v genetické informaci nám přinesl mnohé, včetně neustále se zvyšující se náročnosti ve zpracování dat. Odečítání sekvencí nukleotidů z gelů je již zastaralé. Vzhledem k obrovskému množství dat, které je schopen jeden sekvenační běh vyprodukovat, se počítače staly v sekvenování naprostou nezbytností, zejména v masivně paralelním sekvenování. Výstupem bývají statisíce sekvencí, pro jejichž zpracování je nezbytný specializovaný analytický software a zkušený bioinformatik.

Biologická data jsou v dnešní době produkována neuvěřitelným tempem. V průměru se velikost databází zdvojnásobí každých 15 měsíců [1]. S tímto obrovským množstvím biologických dat je mnohdy pro výzkumné pracovníky složité získat přístup ke správným nástrojům pro jejich zpracování. Roste povědomí o matematické povaze mnohých biologických procesů a výpočetní a statistické modely nacházejí v biologii významné uplatnění.

Tato práce se snaží nabídnout flexibilní veřejně dostupný balíček nástrojů v jazyce R, který bude sloužit k předzpracování a základní analýze surových sekvenačních dat. Tyto nástroje budou následně otestovány na datech získaných sekvenováním na sekvenátoru Roche 454. Balíček funkcí bude doplněn přehlednou dokumentací a nahrán na GitHub, odkud bude každému uživateli volně k dispozici k nainstalování do R.

Cílem práce je také čtenáři přiblížit vybrané sekvenační technologie od počátků sekvenování až po nejmodernější platformy. Dále bude vysvětlena důležitost genotypizace ve výzkumu infekce virem HIV a nutnost využití zvířecích modelů k lepšímu pochopení reakce lidského imunitního systému na tuto infekci.

V poslední části práce bude přistoupeno otestování funkčnosti navrženého balíčku nástrojů na souboru surových sekvenačních dat získaných MHC a KIR genotypizací u makaků ve výzkumu infekce virem HIV.

1 Sekvenační technologie

Metody sekvenování DNA mají za úkol stanovit primární strukturu, tzn. pořadí nukleotidů v molekulách DNA. Z této znalosti primární struktury DNA můžeme odvodit informace o tvorbě proteinů nebo například o genetických mutacích, které jsou příčinou vzniku různých geneticky podmíněných chorob. Následující kapitoly se budou zaobírat přehledem nejpoužívanějších sekvenačních metod od historie až po současnost.

1.1 First generation sekvenování

1.1.1 Maxam-Gilbertovo sekvenování

Tato metoda byla publikována už roku 1976 Allanem Maxamem a Walterem Gilbertem. Je založená na chemickém rozštěpení sekvenovaného úseku molekuly DNA na místech, kde se vyskytuje určitá báze [2].

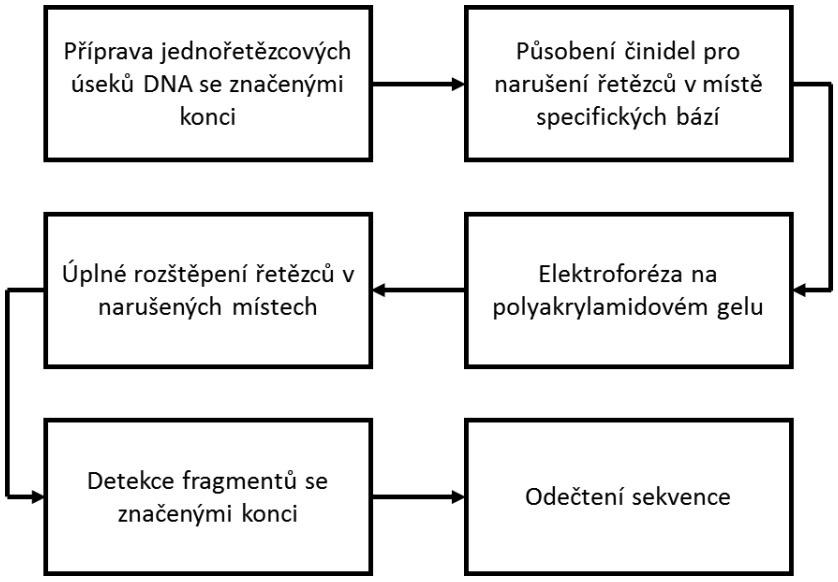
Nejprve je nutno připravit identické jednořetězcové úseky sekvenované DNA značené na jednom konci radioaktivní značkou. Následně takto značenou DNA rozdělíme do čtyř vzorků. Na každém vzorku bude probíhat jiná reakce. Fragmenty jsou vystaveny činidlům, která odštěpí příslušné báze od deoxyribózy. Činidla znázorňuje Tabulka 1.

Tabulka 1: Činidla používaná pro narušení sekvence na příslušných bázích [3]

G	dimetylsulfát
A + G	piperidin
C + T	hydrazin
C	hydrazin + NaCl
A > C	hydroxid sodný

Odhalený sacharid je potom slabým místem v řetězci DNA, takže se při působení vysoké teploty a piperidinu úplně rozštěpí.

Samotná sekvence DNA se vyhodnocuje z elektroforézy na polyakrylamidovém gelu, kde jsou vedle sebe naneseny produkty všech čtyř štěpných reakcí. Následně detekujeme pouze fragmenty, které nesou radioaktivně označený konec. Z poloh pruhů v různých drahách elektroforézy poté stanovíme sekvenci [3]. Kroky této sekvenační metody znázorňuje Obrázek 1.



Obrázek 1: Workflow Maxam-Gilbertova sekvenování.

1.1.2 Sangerovo sekvenování

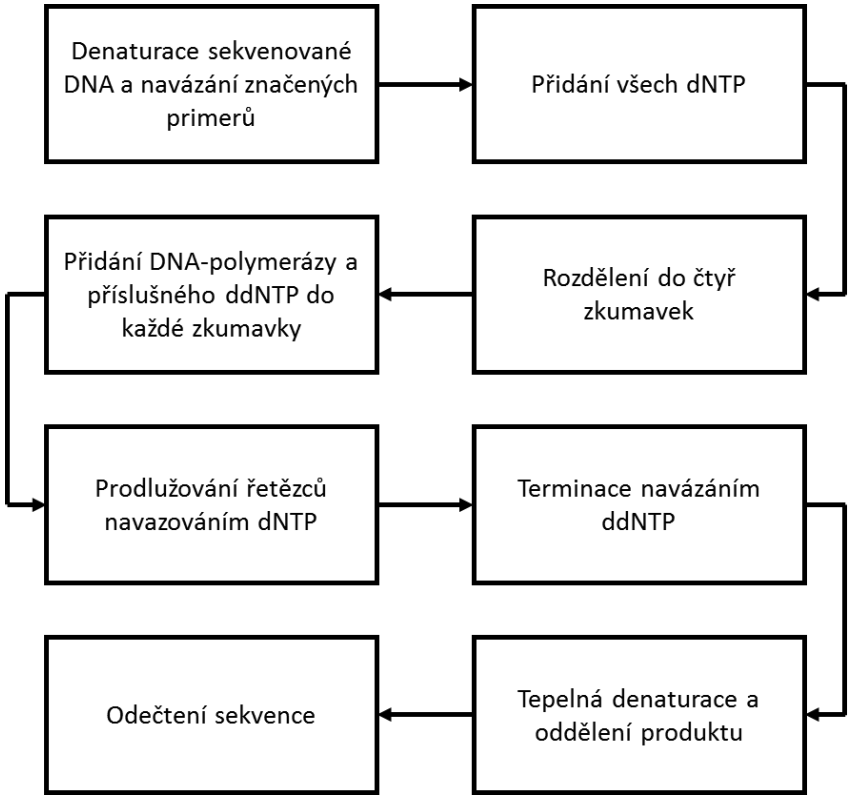
Tato metoda je založena na syntéze komplementárního řetězce DNA s pomocí enzymu DNA–polymerázy. Sekvenovaná DNA slouží jako matrice. Proto tato metoda bývá někdy označována jako enzymová.

Nejprve musí dojít k navázání značeného primeru na úsek jednořetězcové DNA, kterou chceme sekvenovat. Pro to, aby DNA-polymeráza syntetizovala komplementární řetězec, je nezbytná volná 3‘–OH skupina, na kterou se naváže další nukleotid. Syntéza tedy probíhá ve směru 5‘ – 3‘[3].

DNA s navázanými primery je poté, podobně jako u předchozí metody, rozdělena do čtyř vzorků. Ke všem vzorkům je přidán nadbytek dNTP (deoxyribonukleotidtrifosfát) ve všech čtyřech verzích (dATP, dTTP, dCTP, dGTP). Tyto sloučeniny fungují jako stavební kameny pro nově syntetizovaný řetězec DNA. Poté je přidána DNA-polymeráza spolu s příslušnými ddNTP (2‘, 3‘-dideoxynukleotidtrifosfát). V každém ze čtyř vzorků bude jiný ddNTP (tzn. ddATP, ddTTP, ddCTP, ddGTP). dNTP se od ddNTP liší tím, že ddNTP postrádá volnou OH skupinu. Z toho vyplývá, že když DNA-polymeráza použije k prodloužení řetězce ddNTP, syntéza se tím ukončí, protože další nukleotid se již nemá kam navázat. Poměr mezi příslušnými dNTP a ddNTP (například dATP ku ddATP) nám určí, jak budou řetězce syntetizované DNA-polymerázou dlouhé. Zpravidla to bývá 100 dNTP na 1 ddNTP [3].

V každé zkumavce se tedy syntetizuje komplementární řetězec k sekvenované DNA. Prodlužování probíhá do té doby, než je do sekvence zařazen terminátor (ddNTP). V každé ze čtyř zkumavek dostaneme tedy řetězce o různých délkách zakončené příslušnou bází. Vzorky jsou poté tepelně denaturovány, aby došlo k oddělení syntetizovaného řetězce DNA se značeným primerem od matricové DNA.

Stejně jako u Maxam-Gilbertova sekvenování používáme pro vyhodnocení sekvence elektroforézu na polyakrylamidovém gelu. Vzorky ze čtyř zkumavek jsou naneseny vedle sebe. Výsledek elektroforézy vizualizujeme například rentgenovým snímkem, pokud byly primery značeny radioaktivně. Víme, že nejkratší úseky doputují nejdále, takže sekvenci DNA čteme odspodu elektroforetogramu. Takže například, pokud první proužek je v řadě ze zkumavky s přidaným ddATP, víme, že první báze bude adenin. Kroky Sangerova sekvenování znázorňuje Obrázek 2. V komerčním použití Sangerovy metody je využívána kapilární elektroforéza s fluorescenčně značenými fragmenty DNA.



Obrázek 2: Workflow Sangerova sekvenování.

1.2 Next generation sekvenování

Po dlouhou dobu nedocházelo v metodách sekvenování k výraznému vývoji. Sangerova metoda, která byla primárně využívána, měla však svá omezení. Tím hlavním byla nutnost využívání gelů k vyhodnocování výsledků, takže nebylo možné masivně paralelní sekvenování. Proto došlo k vývoji technologií, které gely nevyužívají [4].

1.2.1 Roche 454 pyrosekvenování

Roku 2005 byl představen vůbec první next generation sekvenovací systém firmou 454 Life Sciences. DNA je fragmentována do úseků o délce asi 500 bází. Na fragmenty jsou poté upevněny specifické adaptéry, které zajistí zafixování fragmentů do speciálních kuliček, kdy každá kulička obsahuje právě 1 fragment. V kuličkách poté probíhá

emulzní polymerázová řetězová reakce k amplifikaci fragmentů. Po dokončení PCR jsou DNA v kuličkách denaturovány a umístěny na optický čip, ze kterého vedou statisíce optických vláken napojených na CCD kameru. Poté již probíhají reakce popsané níže. Výstup může dosahovat až 1 Gb za den, což je 1/3 lidského genomu [5]. To bylo v dobách bez masivně paralelního sekvenování jen těžko představitelné.

Metoda detekce pyrofosfátu byla popsána již roku 1985 a poprvé byla k sekvenování DNA použita roku 1988 [4]. Při enzymatické syntéze DNA se z dNTP uvolňuje pyrofosfát PPi. Tento pyrofosfát můžeme kaskádou enzymatických reakcí převést na viditelné světlo.

Na začátku máme jednořetězcovou DNA, která je produktem PCR a slouží jako templát. Toto vlákno inkubujeme spolu se sloučeninami nezbytnými pro pyrosekvenování (viz Tabulka 2).

Potom budeme postupně přidávat jednotlivě všechny 4 dNTP. Výjimka je pouze u dATP, protože do jedné z nejdůležitějších reakcí pyrosekvenování vstupuje právě ATP, takže nemůžeme dATP použít jako zdroj adeninu pro syntézu řetězce DNA. Z tohoto důvodu používáme místo standartního dATP 2´-deoxyadenozin-5´-α-tio-trifosfát, který reaguje s DNA-polymerázou, ale již nereaguje s luciferázou a luciferinem [3].

Tabulka 2 Přehled sloučenin a enzymů nezbytných pro pyrosekvenování.

Sloučenina	Funkce
DNA- polymeráza	Enzym katalyzující polymeraci řetězce DNA.
Apyráza	Degraduje nevyužité dNTP, aby mohl cyklus reakcí pokračovat bez chyb.
ATP sulfuryláza	Enzym katalyzující reakci APS s PPi.
Adenozin-5'-fosfosulfát (APS)	Reakcí pyrofosfátu s PPi vzniká adenosintrifosfát ATP.
Luciferáza	Enzym katalyzující reakci ATP s luciferinem.
Luciferin	Reakcí s ATP vzniká oxyluciferin a viditelné světlo, které detekujeme.

Začneme například dTTP. Pokud je na sekvenovaném vlákně DNA za navázaným primerem přítomen adenin, DNA-polymeráza katalyzuje navázání adeninu za primer. Vedlejším produktem této reakce je pyrofosfát (a). Pyrofosfát je nyní reakcí s adenozin 5‘-fosfosulfátem (APS), katalyzovanou ATP sulfurylázou, převeden na adenozintrifosfát (b). ATP je nezbytné pro reakci katalyzovanou luciferázou, která převede přítomný luciferin na oxyluciferin. Tato reakce zároveň vytvoří světelný záblesk, který detekujeme (c). Pokud aktuální dNTP v reakci není komplementární k nukleotidu na sekvenovaném vláknu DNA, je degradován apyrázou (d) [3].

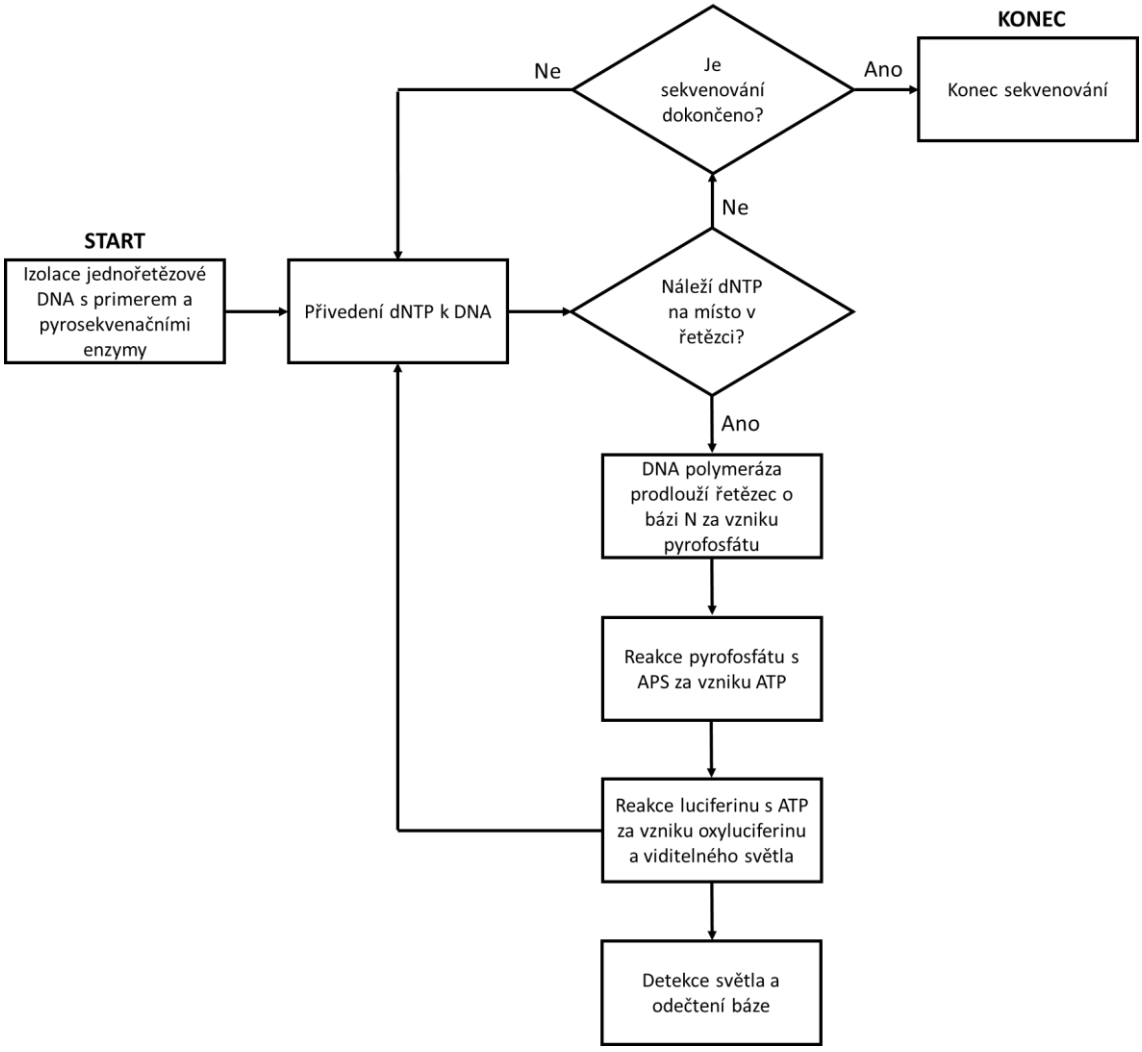
a $DNA_n + dNTP \xrightarrow{DNA-polymeráza} DNA_{n+1} + PPi$

b $PPi + APS \xrightarrow{ATP\ sulfuryláza} ATP + SO_4^{2-}$

c $ATP + luciferin + O_2 \xrightarrow{luciferáza} AMP + PPi + oxyluciferin + CO_2 +$ ***světlo***

d $dNTP \xrightarrow{apyráza} dNMP$

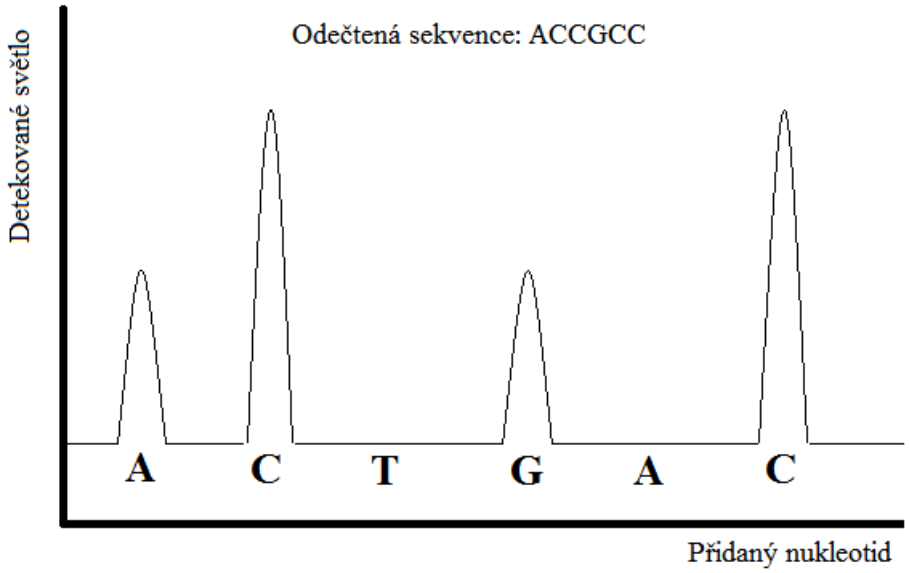
Reakci je opakována, dokud není osekvenované celé vlákno. Rychlost odečítání je přibližně 1 báze za minutu. Workflow pyrosekvenování znázorňuje Obrázek 3.



Obrázek 3: Workflow pyrosekvenování.

Výsledná sekvence je odečítána z pyrogramu (viz Obrázek 4), kde na se na ose x střídají báze A, C, G a T, na ose y je znázorněno množství odečteného světla. Nepřítomnost vrcholu grafu značí, že se na dané pozici určitá báze nevyskytuje. Pokud je za sebou více stejných bází, je detekován vrchol, který je úměrně vysoký. Například, pokud se navážou dva tyminy za sebou, je detekován vrchol, který je přibližně dvakrát

tak vysoký, než vrchol jednoho tyminu. Zde však dochází k tzv. chybovosti na homopolymerech, která je u Roche 454 dobře zdokumentována. Homopolymery jsou úseky tří a více stejných bází. Chybovost je způsobena odchylkami při snímání akumulovaného světla [6].



Obrázek 4: Nákres pyrogramu.

1.2.2 Illumina

Illumina je také metoda podporující masivně paralelní sekvenování. Jde v současné době o nejpoužívanější metodu z next generation sekvenovacích metod. Jednořetězcové úseky DNA jsou umístěny na amplifikační destičku (flow cell), kde jsou vlákna DNA amplifikována v těsné blízkosti u sebe můstkovou amplifikací. Vznikají tak husté shluky miliónů identických jednořetězcových vláken DNA připevněných na amplifikační destičku.

Samotné sekvenování probíhá po navázání primeru a následné syntéze komplementárního řetězce k řetězci fixovanému na amplifikační destičce. Jednotlivé dNTP jsou označeny fluorescenčním barvivem. Tyto fluorescenčně značené nukleotidy jsou chemicky chráněné na 3’ hydroxylovém konci, což zabraňuje navázání více než jednoho nukleotidu v cyklu. Po navázání dNTP a osvícení zdrojem záření fluorescenční značka emituje záření o charakteristické vlnové délce pro každou bázi. Toto emitované záření je sejmuto kamerou a dojde k enzymatickému odtržení chemicky chráněného 3’hydroxylového konce a fluorescenční značky, aby bylo možno pokračovat v dalším cyklu sekvenace. Toto postupné snímání a odtrhávání fluorescenčních značek je velice žádoucí z hlediska sekvenování homopolymerů (opakujících se sekvencí jednoho nukleotidu), se kterým má mnoho jiných sekvenačních platforem problémy. Illumina má oproti jiným next generation sekvenačním platformám ještě tu výhodu, že její výkon je 10-100krát vyšší. Problém Illuminy může spočívat v tom, že kumulace náhodně

enzymaticky neodtržených fluorescenčních značek a chráněných hydroxylových skupin vede v pozdějších fázích sekvenování ke zvyšování šumu, a v pozdějších sekvenačních cyklech může docházet k substitučním chybám [7].

1.2.3 SOLiD

Zatímco obě předchozí zmíněné metody sekvenovaly syntézou („sequencing by synthesis“), technologie SOLiD k určení sekvence DNA využívá metodu ligace („sequencing by ligation“) [8].

Na konce fragmentu DNA jsou nejprve ligovány specifické adaptory, které slouží k hybridizací vlákna DNA ke kuličce, na které bude probíhat emulzní PCR. Jsou dva způsoby, kterými se toto provádí. Buď se používá tzv. fragmentová knihovna, kdy jsou na daný fragment z obou konců připojeny adaptory, nebo tzv. mate-paired knihovna, kdy namísto jednoho fragmentu budeme mít dva se známou sekvencí mezi nimi s tím, že adaptory se ligují na volné konce těchto fragmentů [9].

Dále následuje amplifikace DNA. Stejně jako u Roche 454 se zde používá emulzní PCR. Kuličky s namnoženou DNA jsou poté kovalentními vazbami umístěny na destičku, na které probíhá samotná sekvenace.

Sekvenace probíhá tak, že na vlákno DNA na kuličce nasedne primer. V roztoku je enzym ligáza a dinukleotidové sondy, které jsou označeny fluorescenční značkou. Sondy kompetetivně nasedají na sekvenovaný řetězec a jsou ligovány k vláknu s primerem. Změří se fluorescence, fluorescenční značka je odstřihnuta a je ligována další sonda. Po požadovaném počtu cyklů se nově vytvořený řetězec odstřihne a na sekvenovaný řetězec je hybridizován nový primer, který je posunutý o jednu bázi ve směru sekvenování. Toto resetování primeru je provedeno celkem pětkrát, což zajistí, že každá báze je přečtena dvakrát, takže je zvýšená přesnost čtení [9].

1.2.4 Ion torrent

Next generation sekvenační metoda Ion torrent využívá polovodičového sekvenátoru, na kterém s pomocí polymerázové řetězové reakce probíhá klonální amplifikace knihovny DNA fragmentů na povrchu mikročástic. Tyto mikročástice s navázanými fragmenty DNA jsou následně uloženy do mikrojamek, které jsou schopny rozeznat malé změny v pH při syntéze komplementárního řetězce DNA. Samotný sekvenační process spočívá v adici nemodifikovaných nukleotidů dATP, dGTP, dCTP a dTTP a následným vymytím neinkorporovaných nukleotidů. Takto se naváže vždy na 3’ konec DNA fragmentu pouze jedna komplementární báze [10].

Při polymerační reakci je vedlejším produktem iont pyrofosfátu, který slabě změní pH v jamce. Tato změna je změřena a převedena na daný nukleotid. V případě, že se na sekvenovaném vláknu DNA nachází více identických nukleotidových bází za sebou, naváže se více komplementárních nukleotidů za sebou, a změna pH je tomu úměrná. Protože Ion Torrent využívá jednoduchý system adice nukleotidů a jejich

vymývání, je tato sekvenační plaforma rychlejší než jiné next generation sekvenační metody. Nevýhodou Ion Torrentu je chybovost na homopolymerech od osmi bazí výše [10].

1.3 Third generation sekvenování

Přechod mezi druhou generací sekvenačních technologií a sekvenováním třetí generace je poněkud méně jasný než přechod mezi první a druhou generací. Přeci jen bylo masivně paralelní sekvenování poměrně velký skok od poměrně složitého a časově náročného vyhodnocování sekvencí z gelů. Ve druhé generaci sekvenování je princip většinou takový, že k sekvenovanému vláknu je přidána reakční směs, poté je vymyta a detekuje se příslušný jev, který určí charakter nukleotidů, znovu je přídána reakční směs a takto se určuje nukleotid za nukleotidem. Ve třetí generaci je snaha o to, aby nemuselo docházet k neustálému přerušování reakcí probíhajících při sekvenování a čtení probíhalo v reálném čase. Dojde tak k podstatnému zrychlení celého procesu a z toho důvodu, že DNA není třeba amplifikovat, mohou být jednotlivá čtení mnohem delší [11].

1.3.1 Pacific Biosciences SMRT

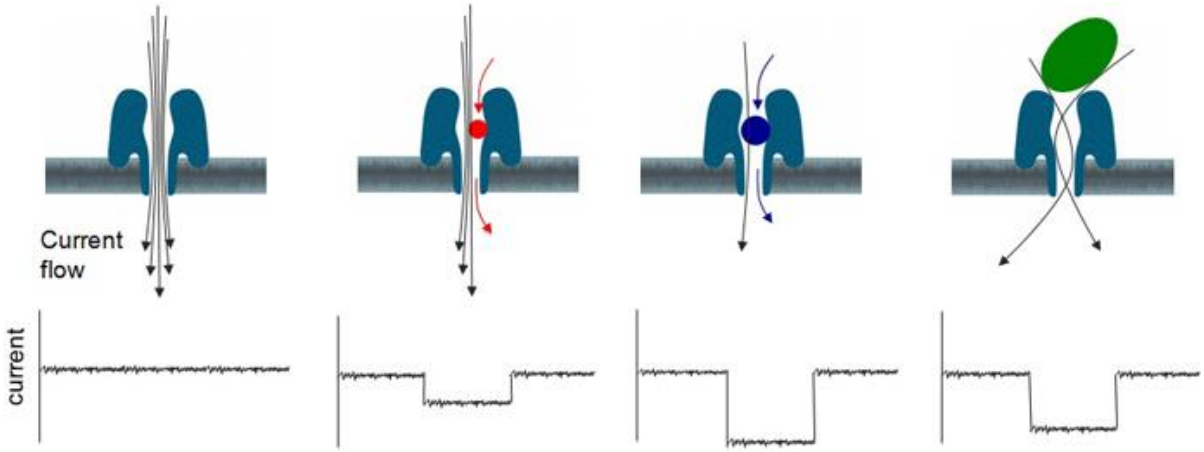
SMRT (single molecule real time) sekvenování je systém, který pozoruje jedinou molekulu polymerázy při tom, jak syntetizuje vlákno DNA, takže přímo využívá rychlosti a efektivity této reakce a nevzniká nutnost vytváření shluků molekul DNA.

Jádrem tohoto sekvenovacího systému je zero-mode waveguide (ZMW) technologie. ZMW jsou maličké nanojamky o velikosti asi 70 nm v průměru a 100 nm v hloubce, v nichž je vložena jediná molekula polymerázy [11]. Těchto jamek mohou být až desetitisíce vedle sebe, takže masivně paralelní sekvenování je samozřejmostí.

V každé jamce DNA polymeráza syntetizuje vlákno komplementární k templátovému vláknu z fluorescenčně značených nukleotidů dATP, dGTP, dCTP a dTTP. Po navázání nukleotidu na DNA fluorescenční značka na konci fosfátové skupiny opouští jamku. I přesto, že kterýkoli fluorescenčně značený nukleotid může jamky jak volně opouštět, tak do nich volně vstupovat, pouze nukleotid inkorporovaný do komplementárního řetězce DNA zůstává delší dobu u dna jamky z důvodu jeho enzymatického zařazení do řetězce. Detektor v dolní části zařízení detekuje vlnovou délku fluorescenční značky tohoto nukleotidu a na základě této vlnové délky zjišťuje, o který nukleotid se jedná. Protože sekvenátor jakkoli nepozastavuje proces přirozené reakce polymerázy, je sekvenování velice rychlé a délka čtení velká. Nevýhodou Pacific Biosciences SMRT je poměrně nízká přesnost sekvenování (chybovost bývá větší než 10 %)[11].

1.3.2 Oxford nanopore

Oxford nanopore je technologie založená na třech molekulách, které byly navrženy tak, aby pracovaly jako systém. Nanopór je konstruovaný z modifikovaného α-hemolysinového póru, který má navázanou exonukleázu na extracelulární straně. Syntetický cyklodextrinový senzor je kovalentně navázán na vnitřním povrchu nanopóru. Celý tento systém je uložen v syntetické lipidové dvojvrstvě, takže když je DNA vložena na povrch nanopóru s exonukleázou a na lipidové dvojvrstě je změněno napětí, exonukleáza je schopna „ustřihnout" individuální nukleotidy z této DNA. Tyto individuální nukleotidy jsou detekovány tím, jak specificky narušují iontový proud probíhající pórem. Průběh reakce znázorňuje Obrázek 5.



Obrázek 5: Průběh narušení iontové proudu probíhajícího nanopórem. Převzato z http://phys.org/news/2014-02-oxford-nanopore-unveils-portable-genome.html.

Jelikož tato metoda k detekci nevyužívá optiku, ale změny elektrického prostředí, je možno dosáhnou velice nízké ceny za detekovanou bázi [11].

2 Techniky sekvenování

Next-generation sekvenačními metodami je možné zároveň sekvenovat statisíce až miliony různých řetězců zároveň, ovšem za cenu toho, že tyto řetězce musí být poměrně krátké. Zároveň, genomy savců jsou velice komplexní a jejich klonování je nesmírně složitá záležitost. Proto je výhodnější a rychlejší, když dlouhé sekvence DNA „rozkrájíme“ na kratší části. Musíme tedy zvolit vhodnou strategii, abychom toto mohli provést.

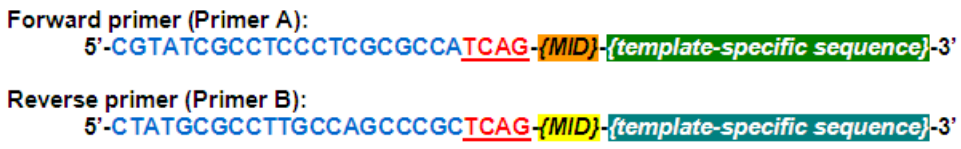
2.1 Shotgun sekvenování

Shotgun sekvenování dostalo svůj název podle náhodného rozptylu výstřelu brokovnice. Sekvenována DNA je náhodně fragmentována, na tyto fragmenty jsou ligovány specifické adaptory a sekvenování pokračuje podle použité technologie. Výsledná čtení jsou následně zarovnána do výsledné sekvence. Toto sekvenování je výhodné pro celogenomové sekvenování.

2.2 Amplikonové sekvenování

Amplikonové sekvenování spočívá v tom, že na části DNA, které jsou sekvenovány, jsou navázány speciální forward a reverse primery, které vymezí sekvenovanou oblast. Tyto úseky DNA jsou poté amplifikovány a osekvenovány. Proto, abychom mohli designovat primery, musíme mít o sekvenované DNA poměrně velkou znalost. Hodí se tak například k vyhledávání jednonukleotidových polymorfismů. Je využíváno takzvaných fúzních primerů (viz Obrázek 6).

2.2.1 Design fúzních primerů



Obrázek 6: Design fúzních primerů (Převzato z [13])

5´- konec primerů je specifickým adaptorem, který složí k hybridizaci vlákna DNA na kuličku v emulzní PCR.

Následuje klíč TCAG, který slouží pro identifikování kontrolních emPCR kuliček od těch, kde je navázán amplikon. Dále slouží sekvenátoru pro kalibraci signálu vzniklého po inkorporaci dané báze [13].

MID je multiplexační identifikátor. Jeho použití není nutné, ale pokud v jednom sekvenačním běhu sekvenujeme DNA z více vzorků, používá se pro rozlišení těchto sekvencí. V anglické literatuře se někdy také místo označení MID používá „barcode“ [13].

Poslední je úsek sekvence nukleotidů komplementární k sekvenovanému vláknu DNA. Správné navržení primeru vyžaduje detailní znalost sekvenované DNA. K optimálnímu navržení této komplementární sekvence se obvykle používá počítačový program, který navrhne vhodnou délku a umístění tohoto primeru. Délka této templátově specifické sekvence bývá obvykle okolo 20-25 bází [13].

Primery jsou dva. Nasedají na fragment DNA z obou směrů. To umožňuje obousměrné sekvenování. Tím, že jsou získány jak dopředné, tak zpětné čtení sekvence, dosáhneme daleko větší přesnosti a správnosti výsledku. Dále je také k přesnosti čtení v Roche 454 systémech velice důležité prostředí okolo sekvenovaného úseku DNA. Může se tedy stát, že úsek DNA, který je obtížně čitelný v dopředném směru, bude ve zpětném čtení čitelný dobře.

2.2.2 Emulsní PCR

Jednořetězcové úseky DNA se hybridizují s pomocí svých adaptorů na speciální kuličku tak, aby na každé kuličce byl navázán právě jeden řetězec DNA. Dále je z kuliček vytvořena olejová emulze, která obsahuje všechny nezbytné složky pro PCR reakci. To vytvoří z kuliček jakési mikroreaktory, v nichž odděleně probíhají polymerázové řetězové reakce. Na konci emPCR, po oddělení olejové složky, budeme mít kvantum kuliček a na každé z nich je až 10 milionkrát amplifikován právě jeden řetězec DNA.

Kuličky jsou poté umístěny na speciální optickou pikotitrační destičku. Do každé jamky na destičce se vejde právě jedna kulička. Jedna kulička tedy znamená jedno čtení. Tyto kuličky jsou poté zasypány dalšími kuličkami, na kterých jsou navázány enzymy nezbytné pro průběh pyrosekvenování [14].

2.2.3 Multiplexace

Plný sekvenační běh má velice často nadměrnou kapacitu pro sekvenování pouze jednoho vzorku DNA. Abychom byli při sekvenování co nejefektivnější, sekvenujeme mnohdy více vzorků v jednom sekvenačním běhu. V tomto případě nám nebudou pro rozlišení sekvencí od jednotlivých vzorků stačit pouze specifické primery nasedající na jednotlivé amplikony, protože tyto primery mohou být pro různé vzorky totožné. Vzniká tedy nutnost použití multiplexačních identifikátorů. Multiplexační identifikátory, dále MIDy, jsou krátké sekvence nukleotidů uměle vložené ve fúzních primerech. Jejich kombinace s primerem nasedajícím na sekvenovanou DNA poté pomáhá jednoznačně určit, které čtení náleží kterému vzorku [13]. Tento proces se nazývá demultiplexace.

3 Genotypizace ve výzkumu infekce virem HIV

HIV je jedním z nejdiskutovanějších světových problémů již od svého objevu roku 1981. Infekce virem HIV je charakterizována akutní virémií. V jednom mililitru plazmy se může vyskytovat až 5 miliónů virových částic. V případě, že člověk nepodstupuje retrovirovou léčbu, nastupuje snižování počtu $CD4^+$ T-lymfocytů a zvyšování koncentrace virových částic v plazmě [15]. Nastupuje tak AIDS (Acquired Immune Deficiency Syndrome). Malá část populace však vykazuje schopnost úspěšně udržovat koncentraci viru na stabilní úrovni i bez terapie, jsou si tedy schopni zachovat dostatečný počet $CD4^+$ T-lymfocytů, nenastupuje u nich AIDS a pravděpodobnost, že infekci předají dál, je menší, než u lidí, kteří tuto přirozenou rezistenci nevykazují [15].

Při studii lidského genomu bylo určeno, že za tuto přirozenou rezistenci proti viru HIV mohou převážně jednonukleotidové polymorfismy na šestém chromozomu, v oblasti hlavního histokompatibilního komplexu (MHC) [15].

3.1 Hlavní histokompatibilní komplex (MHC)

Hlavní histokompatibilní komplex je skupina genů, která zajišťuje adaptivní autoimunitní reakci buněk. Jeho poznání je proto nezbytné pro výzkum infekčních nemocí, vývoj vakcín a pro transplantační medicínu. U člověka je tato oblast genů označována jako „human leukocyte antigens“ (HLA) [16].

Hlavní funkcí molekul hlavního histokompatibilního komplexu je vychytávat fragmenty proteinů, které vznikly například patogenním působením virů, z vnitřního prostředí buňky a vystavit je na buněčný povrch, kde budou v dosahu imunitního systému. Výsledkem bývá zabití virem infikované buňky, nebo aktivace makrofágů pro zastavení bakteriální infekce a aktivace lymfocytů pro vytvoření protilátek proti toxickým produktům [17].

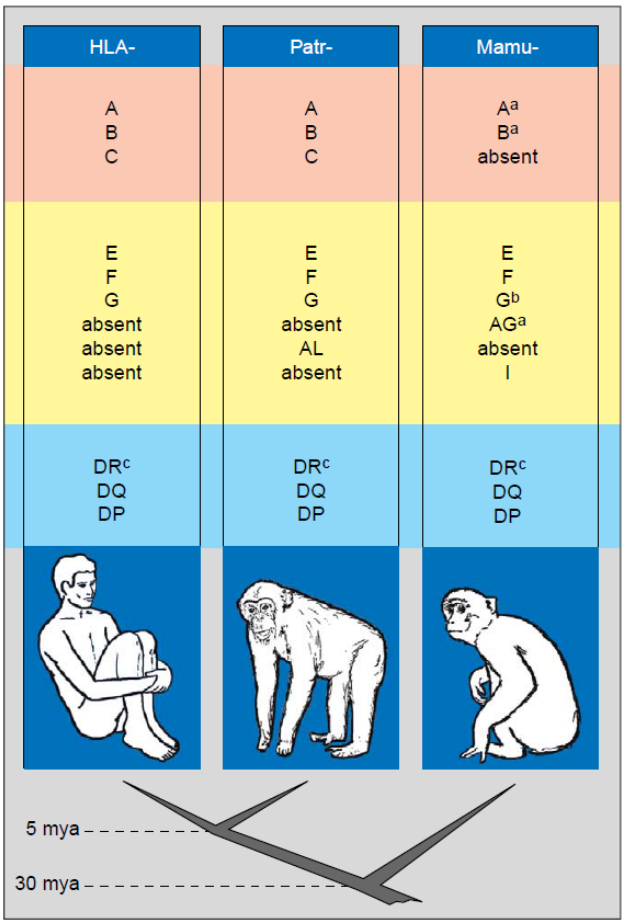
Hlavní histokompatibilní komplex obsahuje dva shluky lokusů, které kódují bílkoviny buněčného povrchu hrající kritickou roli při adaptivní imunitní reakci. Normálně jsou na těchto povrchových bílkovinách vystaveny peptidy odvozené z těchto bílkovin, avšak v případě infekce mohou tyto vystavené peptidy být odvozeny od virů, nebo jiných intracelulárních parazitů, které se v zasažené buňce nacházejí. CD8+ cytotoxické T-lymfocyty tyto peptidy odvozené z patogenů, vystavené na MHC I molekulách, rozpoznají jako cizí, což vede k lýze zasažené buňky. Některé viry se snaží vyhnout rozpoznání imunitním systémem tak, že interferují s transkripcí MHC I, a tak zabraňují expresi genu. Například HIV-I *Nef* protein selektivně inhibuje tvorbu povrchových HLA-A a HLA-B molekul, které jsou kritické pro rozpoznání CD8+ T-lymfocyty. *Nef* proteiny modulují expresi MHC I i u subhumánních primátů zasažených infekcí SIV (symian immunodeficiency virus)[18].

Zatímco MHC I molekuly se vyskytují téměř na všech buňkách, MHC II molekuly jsou přítomny pouze u specializovaných buněk imunitního systému. Tyto molekuly mají za úkol zachytávat a vystavovat peptidy vzniklé extracelulárním působením bakterií a virů. Tyto peptidy jsou vystavovány pro CD4+ T-lymfocyty. MHC II molekuly jsou klíčové při tvorbě protilátek, ale také napomáhají CD8+ T-lymfocytům [18].

3.1.1 Využití primátů ve výzkumu infekce virem HIV v souvislosti s MHC

MHC různých druhů primátů má společnou řadu vysoce příbuzných lokusů. Tyto lokusy spojuje fakt, že obsahují velké množství polymorfismů především v oblasti kódující oblast, kam molekuly buněčného povrchu váží peptidy. Různé alotypy MHC mohou zapříčiňovat buď susceptibilitu, nebo rezistenci k různých infekcím. Z toho důvodu mohou tyto jednonukleotidové změny na HLA zapříčiňovat selektivní výhodu proti různým infekčním onemocnění včetně HIV [18]. Tato variabilita na MHC také zapřičiňuje, že se snižuje pravděpodobnost toho, že by celá populace mohla být vyhlazena jedním konkrétním patogenem.

Nejbližším žijícím příbuzným člověka jsou šimpanzi, se kterými sdílíme přibližně 98,6% genetickou podobnost. Rozeznáváme dva zástupce šimpanzů: šimpanz učenlivý (*Pan troglodytes*) a šimpanz bonobo (*Pan paniscus*), přičemž se oba nacházejí na africkém kontinentu. Ve výzkumu HIV se však ve většině případů využívá zástupců rodu makaků, konkrétně makak rhéský (*Macaca mulatta*) Makakové obývají nejširší geografickou oblast ze všech druhů subhumánních primátů od východního Afghánistánu po západ Číny. Pro výzkum se většinou využívají makakové pocházející z Indie a Číny [18]. Obrázek 7 znázorňuje příbuznost MHC u lidí, šímpanzů a makaků.



Obrázek 7: Distribuce ortologních oblastí MHC u lidí (HLA), šimpanzů (Patr-) a makaků (Mamu-), převzato z [18]

Přenos patogenu z jednoho živočišného druhu na jiný se nazývá zoonóza. HIV dělíme na dvě třídy: HIV-1, HIV-2. Více než 30 druhů afrických kočkodanovitých primátů je přirozeně infikováno různými kmeny SIV. Infekce těchto od narození zasažených primátů je charakterizována vysokými hladinami viru v krevní plazmě, přičemž se ale zřídkakdy vyskytují negativní projevy s touto infekcí spojené. Mangabej kouřový (*Cercocebus atys*), který je pravděpodobně zdrojem infekce kmenem HIV-2 u lidí, však po delší době příznaky infekce vykazuje [18]. Kmeny SIV využívané ve výzkumu infekce virem HIV pocházejí právě od těchto benigně infikovaným magabejů kouřových [19]. Oproti většině afrických kočkodanovitých primátů, kteří příznaky AIDS nevykazují, u kočkodanovitých původem z Asie (převážně makaků) po experimentálním infikováním SIV nastupují příznaky podobné AIDS do dvou let [19].

Z tohoto důvodu jsou tito asijští kočkodanovití primáti, makakové, nejlepším živočišným modelem pro studium infekce virem HIV u lidí. Další výhodou makaků pro studium HIV a AIDS je fakt, že reakce jejich imunitního systému na SIV je velice podobná reakci lidského imunitního systému na SIV, což je činí perfektním kandidátem pro studium reakce lidského imunitního systému na HIV-1, který je blízkým příbuzným SIV [18]. Makakové infikováni SIV podstupují podobnou ztrátu CD4+ T-lymfocytů, jakou zažívají lidé infikováni HIV. Nevýhodou je že, znalost hlavního

histokompatibilního komplexu makaků zaostává za znalostí lidského a organizace MHC makaků je odlišná.

Vhledem k velkému počtu vakcín, které je nutno nejprve testovat na zvířatech je využití makaků nezbytné. Pro výzkum se používají 3 typy makaků: makak rhesus, makak jávský a makak vepří. Makak rhesus má dva podtypy: čínský a indický. Při aplikaci výsledků z výzkumu na makacích na člověka je třeba brát v potaz, že zkoumané geny (např. MHC) se liší [20].

U člověka i u makaka se MHC nachází na šestém chromozomu v navzájem ortologních regionech [21]. Hlavní rozdíl mezi hlavním histokompatibilním komplexem člověka a makaka rhesa (*Macaca mulata*) se nachází v jeho velikosti. Zatímco velikost MHC člověka je asi jenom 3,7 Mb, velikost MHC makaka dosahuje až k 5,3 Mb. Hlavní část tohoto velikostního rozdílu se nachází v třídách I A a B regionu MHC. V ostatních částech jsou MHC člověka a makaka téměř identické.

Člověk má oproti makakovi navíc 4 geny, makak oproti člověku 6 genů. Dalších 8 genů je společných pro člověka i makaka, ale jsou alternovány různými začátky a konci kódování proteinů. Zbytek asi 140 genů se zdá být společný pro oba MHC [22].

3.2 Killer cell immunoglobulin-like receptors (KIR)

Ve výzkumu infekce virem HIV a boji našeho imunitního systému proti ní jsou velice důležité tzv. NK buňky (natural killer). Tyto buňky řadíme mezi lymfocyty. Jejich úkolem je rozeznat nádorové buňky, buňky napadené virovými infekcemi nebo obecně buňky, které vykazují stres.

Pro funkčnost těchto „přirozených zabijáků“ jsou nezbytné aktivační a inhibiční proteinové receptory, které jsou schopny rozeznat přítomnost abnormalit. Jedním typem těchto receptorů jsou právě proteiny kódované KIR geny. Stejně jako geny kódující MHC vykazují geny kódující KIR velkou genetickou variabilitu ve formě velkého množství polymorfismů, proto se na ně zaměřuje množství studií zkoumajících reakce organismu na vir HIV [23].

3.3 Biologické databáze ve výzkumu HIV

Biologické databáze jsou nezbytné pro téměř každý biologický výzkum. Sdružují data z literatury, vědeckých experimentů, masivně paralelních sekvenačních metod a počítačové analýzy. Obsahují informace z genomiky, proteomiky, fylogenetiky aj. Biologické databáze dělíme na dva typy, primární a sekundární, které se liší svou strukturou. Primární databáze většinou sdružují data určitého typu a spravují je ve svém archivu. Pravidelně nahrávají nová data a aktualizují stará data. Sekundární databáze jsou databáze, které sdružují data právě z primárních databází. Často tyto data různě zpracovávají a analyzují, což vede k získávání nových výsledků. Mezi takovéto sekundární databáze patří například NCBI (National Center for Biotechnology

Information), EMBL-EBI (European Molecular Biology Laboratory – European Bioinformatics Institute) a DDBJ (DNA Data Bank of Japan).

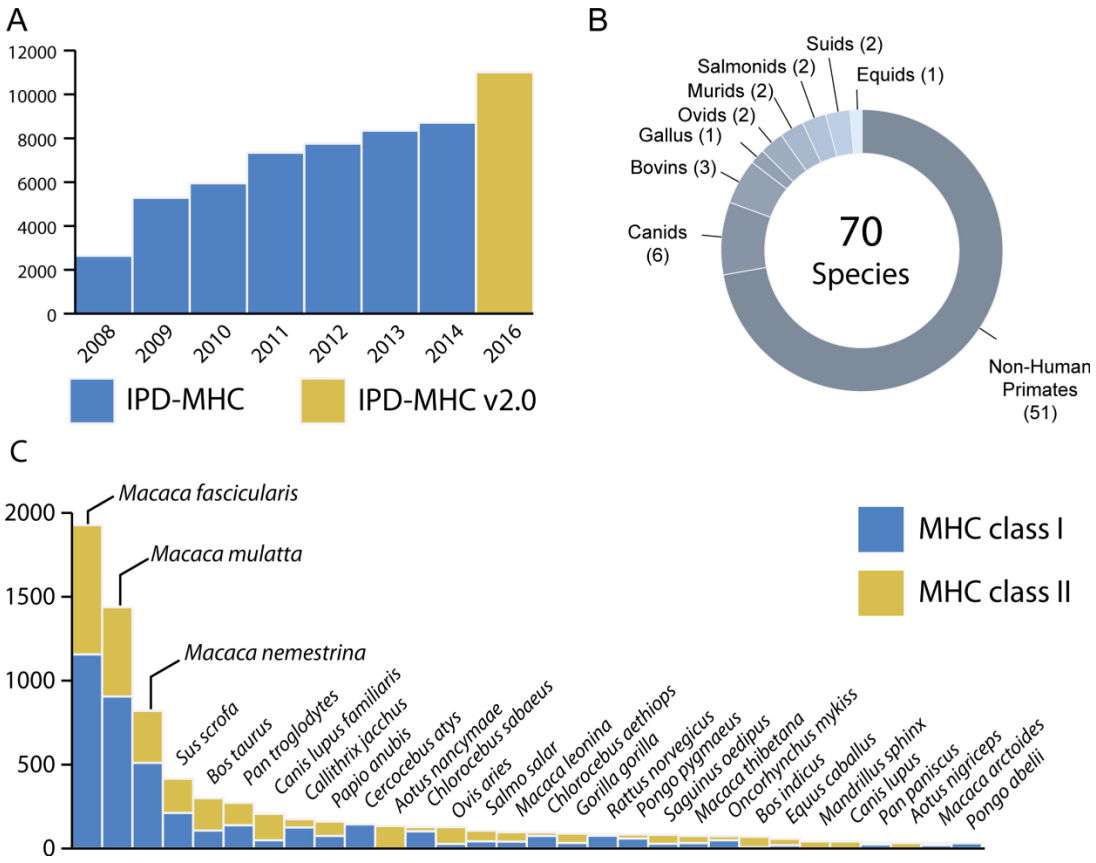
Pro potřeby práce byla zvolena databáze IPD (Immuno Polymorphism Database) dostupná přes webové stránky a ftp server EMBL-EBI. EMBL-EBI je instituce založená roku 1992, která je centrem bioinformatického výzkumu v Evropě. Poskytuje širokou základnu databází a nástrojů bioinformatiky a výpočetní biologie, mezi které patří například BLAST, Clustal Omega a mnoho dalších [24].

3.3.1 Immuno polymorphism database (IPD)

IPD je sada databází specializovaných na polymorfismy genů imunitního systému. V současnosti obsahuje celkem čtyři databáze. IPD-KIR obsahuje sekvence alel lidských killer cell immunoblobuline-like receptorů. IPD-MHC zahrnuje sekvence hlavního histokompatibilního komplexu různých druhů. IPD-HPA obsahuje alloantigeny exprimované na krevních destičkách a nakonec IPD-ESTDAB (European Searchable Tumour Cell-Line Database), která poskytuje přístup do databáze imunologicky charakterizovaných buněčných linií melanomu [25].

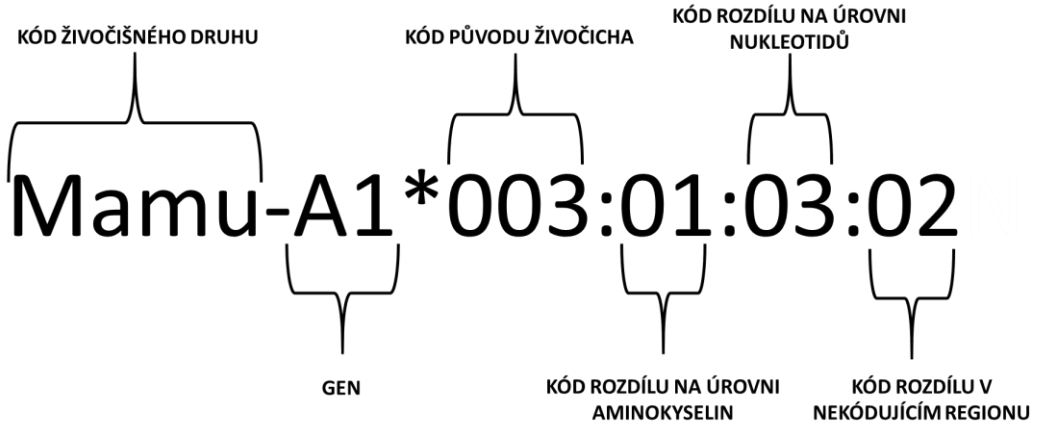
Sekce IPD s podobnými daty (MHC a KIR) sdílejí stejnou strukturu databáze, což je velká výhoda pro návrh a implementaci univerzálních nástrojů pro získávání dat z těchto databází. IPD také obsahuje množství nástrojů pro práci s těmito daty, které jsou buď inkorporovány do existujících nástrojů poskytovaných EMBL-EBI, nebo jsou vytvořeny speciálně pro IPD.

IPD obsahuje sekvence hlavního histokompatibilního komplexu celkem 70 organismů, z čehož 51 jsou různé druhy subhumánních primátů (viz Obrázek 8).



Obrázek 8: Skladba IPD-MHC databáze, převzato z [25]

Dříve byla nomenklatura jednotlivých alel MHC jiných než lidských neformálně stanovována individuálními výzkumnými týmy, které nové sekvence do databáze přidávaly, nebo formálně komisemi založenými ISAG (International Society for Animal Genetics). V současnosti je nomenklatura MHC spravována CMHCNC (Comparative MHC Nomenclature Committee), která je podporována ISAG a VIC (Veterinary Immunology Committee) [25]. Způsob pojmenovávání jednotlivých alel znázorňuje Obrázek 9.



Obrázek 9: Nomenklatura alel MHC

4 Balíček nástrojů GenDP pro zpracování surových sekvenačních dat

Pro rychlé a efektivní zpracování surových sekvenačních dat byl programovacím jazyku R vytvořen balíček funkcí, který byl následně nahrán na github a je v R volně dostupný pod příkazem `install_github("janmatula/GenDP")`. Následující kapitoly se budou zabývat popisem jednotlivých nástrojů v balíčku a stručně bude zmíněna role programovacího jazyka R a projektu Bioconductor v bioinformatice.

4.1 Programovací jazyk R

R je vysokoúrovňový programovací jazyk a prostředí pro statistické výpočty a vizualizaci dat. Jde o GNU projekt, který je podobný jazyku S. Mezi jazyky S a R je mnoho podstatných rozdílů, ale kód napsaný v S půjde bez změn spustit i v R. R poskytuje prostředky k mnoha statistickým (lineární a nelineární modelování, analýza, klasifikace) a grafickým technikám a je jednoduše rozšiřitelný. R představuje ideální volbu pro prototypování nových analytických metod. I když jsou tyto metody v jazyce R většinou pomalé, pokud se ukáže, že má daná metoda budoucnost, může být později reimplementována.

R také poskytuje zavedený systém tvorby balíčků softwarových komponent a dokumentace. To je obrovskou výhodou pro tvorbu, testování a hlavně distribuci softwaru. Jádrem jazyka R je CRAN (Comprehensive R Archive Network), který obsahuje tisíce vzájemně kompatibilních balíčků, které řeší široké spektrum statistických, matematických a vizualizačních problémů. Všechny balíčky v CRAN mohou být staženy jakou open source software. R nabízí komplexní sadu funkcí a balíčků pro přístup k databázím a webovým zdrojům (například prostřednictvím http). Další důležitou vlastností R jsou propracované možnosti vizualizace dat. Dalším, a možná nejdůležitějším aspektem R, je aktivní komunita vývojářů, která se neustále snaží zdokonalovat staré a vyvíjet nové nástroje a funkce [26].

4.1.1 R Bioconductor

Bioconductor je projekt pro kolaborativní tvorbu sofwaru pro matematickou biologii a bioinformatiku. Mezi hlavní cíle projektu patří zjednodušení přístupu výzkumníku k nástrojům, které mohou přispět jejich výzkumu a dále snaha o dosáhnutí možnosti vzdáleně reprodukovat výsledky výzkumů [26].

4.2 Demultiplexace

Prvním krokem, kterým surová sekvenační data projdou je demultiplexace. Princip využití multiplexačních identifikátorů v amplikonovém sekvenování byl popsán

v kapitole 2.2, která se zabývá amplikonovým sekvenováním. Pro demultiplexaci surových sekvenačních dat byla navržena funkce `demultiplex`.

První úkon, který funkce `demultiplex` vykonává, je převod MIDů na komplementární a zleva doprava převrácenou sekvenci nukleotidů s pomocí funkce `reverseComplement`. Dále jsou z původních MIDů a těchto transformovaných MIDů vytvořeny regulární výrazy, jako v následujícím příkladu.

```
"^ACTGACTGAC.+ACTGACTGAC$"
```

Regulární výrazy jsou slova, která definují určitou strukturu textového řetězce. V tomto případě jsou použity pro identifikaci sekvencí začínajících forward MIDem a zakončených reverzním komplementem reverse MIDu a naopak, sekvence začínající reverse MIDem a končící revezním komplementem forward MIDu. Regulární výrazy jsou konstruovány s pomocí speciálních znaků, které mají vždy určitý význam. Význam znaků využitých při konstrukci regulárních výrazů pro demultiplexaci znázorňuje Tabulka 3.

Tabulka 3: Znaky v regulárních výrazech

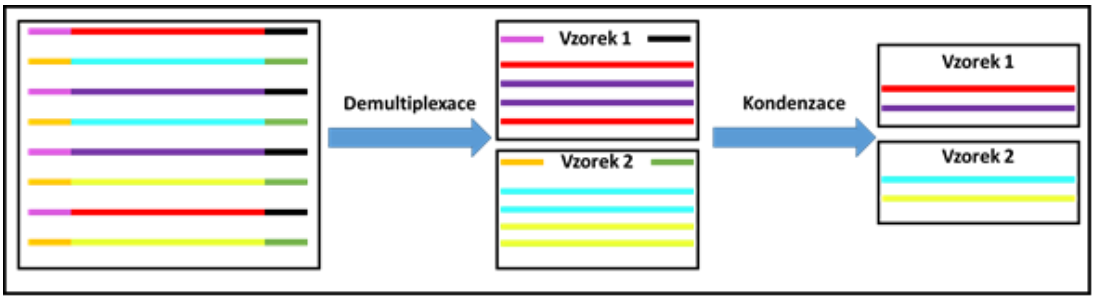
Znak	Význam
.+	nahrazuje sekvenci jakýchkoli znaků, která má délku větší než 1
^	kotva, která značí začátek řetězce
$	kotva, která značí konec řetězce

S pomocí funkce `grep` se soubor surových sekvencí ve formátu FASTA prohledá a naleznou se shody. Z těchto shod se poté vystaví nový FASTA soubor pro daný vzorek, který je následně uložen do pracovní složky R.

4.3 Kondenzace

Demultiplexované FASTA soubory obsahují redundantní informaci v podobě opakujících se čtení těch samých sekvencí. Pro zbavení se těchto nadbytečných sekvencí byla vytvořena funkce `condense`. Funkce najde všechny unikátní sekvence nukleotidů ve FASTA souboru a počet výskytů každé z nich. Unikátní sekvence jsou seřazeny podle jejich počtu a uloženy do nového FASTA souboru. Původní počty sekvencí nejsou ztraceny, ale jsou uloženy do hlaviček sekvencí, kde se k nim může přistupovat.

Volitelným vstupem funkce je práh, který je implicitně nastaven na hodnotu 10. Práh zajišťuje zanedbání všech sekvencí s počtem výskytu nižším než jeho hodnota. Používá se, protože u sekvencí, jejichž počet byl v původním souboru velmi malý, je velice pravděpodobné, že vznikly chybou při sekvenování, a tudíž nemají žádnou vypovídající hodnotu. Schéma kondenzace a demultiplexace znázorňuje Obrázek 10.



Obrázek 10: Schéma demultiplexace a kondenzace

4.4 Stažení referenční databáze

Po sekvenování nových sekvencí DNA je žádoucí tyto nové sekvence porovnat se sekvencemi již existujícími a popsanými. Jde o jednu z nejlepších a nejjednodušších cest, jak přiřadit neznámé sekvenci DNA její funkci [27]. To vyžaduje mnohem více než pouhé porovnání textových řetězců a různé algoritmy berou v potaz tzv. biologicky signifikantní shodu. Dále je důležitá volba vhodné databáze.

Pro usnadnění stahování FASTA souborů referenční databáze byl v jazyce R vytvořen nástroj pro stahování z ftp serveru EMBL-EBI (European Bioinformatics Institute). Funkce má název `downloadDb`. Pro stažení všech FASTA souborů nukleotidových sekvencí na daném uložišti funkce vyžaduje pouze URL. Funkce z HTML kódu stránky sestaví seznam adres k jednotlivým souborům, které následně stáhne.

Funkce obsahuje několik volitelných vstupů. Uživatel si může zvolit adresář, do kterého se databáze uloží. Dále si může uživatel zvolit, zda chce stáhnout nukleotidové, nebo aminokyselinové sekvence. Posledním vstupem je možnost sloučit všechny databázové soubory do jednoho FASTA souboru, který může uživatel pojmenovat.

Pro uživatele je možnost stažení databáze přímo v jazyce R příjemná, protože není třeba žádných dalších nástrojů a databáze je ihned připravena k použití. Oproti manuálnímu stahování souborů má funkce očividnou výhodu v rychlosti.

4.5 Vyhledávání v referenční databázi

Z velkého množství různých algoritmů byl v práci zvolen algoritmus SSAHA, který je svým přístupem k vyhledávání v databázích DNA zcela unikátní.

4.5.1 SSAHA

SSAHA (Sequence Search and Alignment by Hashing Algorithm) je rychlá metoda pro hledání ve velkých databázích DNA. Experimenty ukázaly, že SSAHA může být třikrát až čtyřikrát rychlejší než BLAST a vyžaduje méně paměti, než metody založené na sufixových stromech [28]. Rychlost algoritmu spočívá především ve skutečnosti, že největší část jeho výpočetní náročnosti náleží zahashování referenční databáze, které je nutno provést pouze jednou. Jakkoli velká databáze sekvencí DNA je metodou

převedena do dvou jednoduchých číselných struktur a seznamu názvů sekvencí. Jakmile jsou tyto struktury vytvořeny, vyhledávání v nich je již velice rychlé.

Hlavním úkolem této metody je najít v databázi $D = \{S_1, S_2, S_3, ..., S_i\}$ výskyt sekvence *Q*. Každá sekvence v *D* je označena indexem *i*. Sekvenci DNA, která je *k* bází dlouhá, označujeme jako *k*-mer. Pozici *k*-meru vůči první bázi v sekvenci S označíme jako *j*. Z toho vyplývá, že pozice každého *k*-meru v *D* může být jednoznačně určena párem indexů *i* a *j* [28]. Pokud je množství stejně dlouhých unikátních sekvencí nukleotidů o délce *k* převedeno na jejich binární reprezentaci, pak lze převedením těchto binárních čísel do desítkové soustavy jednoznačně označit každou jedinečnou sekvenci o dané délce *k* integerem od 0 do 4^k-1 (viz Tabulka 4). Této vlastnosti binární reprezentace využíváme k vytvoření hashovací tabulky z databáze a následnému vyhledávání v ní.

Tabulka 4: Konstrukce k-merů, k=3

K-mer	Binární reprezentace k-meru	Integer
aaa	000000	0
aac	000001	1
aag	000010	2
aat	000011	3
aca	000100	4
acc	000101	5
acg	000110	6
act	000111	7
…	…	…
ttt	111111	63

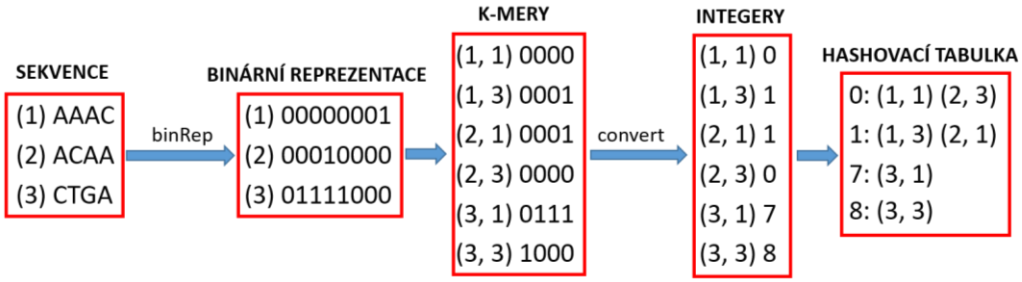
4.5.2 Vystavění hashovací tabulky

Funkce `buildHashTable` zajišťuje vystavění hashovací tabulky z FASTA souboru referenční databáze. Nejprve je databáze načtena do struktury *DNAStringSet* a veškeré znaky, které nezastupují konkrétní báze, jsou převedeny na adenin (například znak N pro neurčenou bázi). Poté jsou tyto sekvence DNA převedeny do binární reprezentace, což zajišťuje funkce `binRep`, která je také součástí balíčku. Funkce si načte sekvence z databáze ve struktuře *DNAStringSet* a převede je na sekvence sestavené z jedniček a nul (Tabulka 5).

Tabulka 5: Binární reprezentace nukleotidů

Báze	Binární reprezentace
adenin	00
cytozin	01
guanin	10
tymin	11

Poté jsou sekvence rozděleny na stejně dlouhé nepřekrývající se úseky (*k*-mery), jejichž délka je vstupem funkce. Tyto úseky jsou přiloženou funkcí `convert` převedeny na 2*k* bitový integer. Z integerů a indexů *i* a *j* odkazujících na *k*-mery se poté vystaví hashovací tabulka (viz Obrázek 11).



Obrázek 11: Vystavění hashovací tabulky z databáze, k=2

Hashovací tabulka má tři části. První částí je seznam hlaviček jednotlivých sekvencí databáze, které jsou zachovávány proto, aby sekvence bylo možno identifikovat. Druhou částí je seznam *k*-merů zakódovaných v podobě integerů. Třetí částí je již seznam indexů *i* a *j*, které ukazují, na které místo v databázi daný *k*-mer náleží. Všechny části jsou v podobě struktury list výstupem funkce. Takto zpracovanou databází lze následně uložit a posléze s ní dále pracovat.

4.5.3 Vyhledávání v hashovací tabulce

Pro vyhledávání v takto vytvořené hashovací tabulce byla navržena funkce `searchHashTable`. Vstupy funkce popisuje Tabulka 6.

Tabulka 6: Vstupy funkce searchHashTable

Vstup	Funkce vstupu
querySequences	Název FASTA souboru obsahujícího hledané sekvence DNA.
hashTable	Hashovací tabulka, která je výstupem funkce `buildHashTable`.
k	Délka *k*-meru u použité hashovací tabulky (implicitně je nastaveno k=10)
results	Počet výsledků pro každou hledanou sekvenci (implicitně nastaveno results = 1).
tolerance	Tolerance v počtu bází pro inzerce a delece (implicitně nastaveno tolerance = 1).

Funkce si nejprve načte vyhledávané DNA sekvence do struktury *DNAStringSet*. Poté dojde k jejich převedení na *k*-mery zakódované 2*k*-bitovým integerem podobně jako u vystavění hashovací tabulky, ale zatímco se u hashovací tabulky *k*-mery nepřekrývají, u převedených hledaných sekvencí ano. Následuje iniciace seznamu matic výsledků. Každá hledaná sekvence má svou matici, která bude obsahovat názvy příslušných nalezených sekvencí z databáze a skóre jejich shody s hledanou sekvencí. Počet řádků v této matici odpovídá vstupu *results*, tedy požadovanému počtu výsledků vyhledávání v databázi. Výsledky budou logicky seřazeny od největšího skóre k nejnižšímu.

Následuje již samotné porovnávání sekvencí vůči zahashované databázi ve *for* cyklu, kde se v každé iteraci vůči databázi porovnává jedna sekvence DNA. Nejprve je třeba vystavět seznam *hitů*, tzn. shod v zahashovaných *k*-merech databáze a hledané sekvence. Jestliže se mezi sekvencí a databází nenajde žádná shoda, je po tomto kroku automaticky zahájena další iterace cyklu. Výsledkem je vektor indexů ukazujících do míst hashovací tabulky, kde se nachází zahashované *k*-mery shodné s hledanou sekvencí DNA. S pomocí těchto indexů jsou z hashovací tabulky vybrány indexy *i* a *j*, a jsou zařazeny do seznamu *M*. Seznam *M* tedy v této fázi algoritmu obsahuje pro každý shodný *k*-mer indexy *i* a *j* ukazující do míst databáze, kde se nacházejí úseky DNA o délce *k* shodné s hledanou sekvencí (index *i* ukazuje na sekvenci v databázi, index *j* na pořadí báze v dané sekvenci, kde shoda začíná).

Nyní se v seznamu indexů *M* přistoupí k výpočtu tzv. *shiftu*. *K*-mery obsažené v seznamu *M* jsou očíslovány od nuly do délky seznamu *M*. Toto číslo nazveme *t*. Následně jsou od indexů *j* shod každého *k*-meru v seznamu *M* odečteny příslušná *t* a toto výsledné číslo je vloženo do příslušného řádku v seznamu *M* jako *shift* (viz rovnice (1)) [28].

(1)

$$M_1 = (i_1, \boldsymbol{j_1} - \mathbf{0}, j_1)$$
$$M_2 = (i_2, \boldsymbol{j_2} - \mathbf{1}, j_2)$$
$$\dots$$
$$M_{\mathrm{n}} = (i_n, \boldsymbol{j_n} - \boldsymbol{t_n}, j_n)$$

Nyní se tedy již seznam *M* sestává ze tří hodnot pro každou shodu: indexů *i* a *j* a vypočteného *shiftu*. V této fázi již mohou být zanedbány zakódované *k*-mery a seznam M je převeden na jednoduchou matici o třech sloupcích (*i*, *shift* a *j*). Tato matice je následně seřazena nejprve podle indexu *i* a následně podle *shiftu*. Nyní se v matici vyhledávají za sebou jdoucí řádky, které mají společné indexy *i* a *shift*.

Vyhledávání těchto posloupností se společným indexem *i* a *shiftem* bylo ve funkci realizováno tak, že každý pro každý unikátní index *i* v *M* je provedena iterace cyklu, při které jsou pro dané *i* nalezeny počty výskytu jednotlivých *shiftů*. Tyto počty výskytu nám určují počet po sobě jdoucích *k*-merů shodných mezi hledanou sekvencí a *i*-tou sekvencí v databázi. Frekvence výskytů *shiftů* jsou následně umocněny dvěma a stává se z nich skóre. Toto umocnění bylo navrženo, aby měly ty nejdelší posloupnosti *k*-merů ve výsledku co největší váhu, zatímco váha osamocených shod zůstane vždy na hodnotě 1.

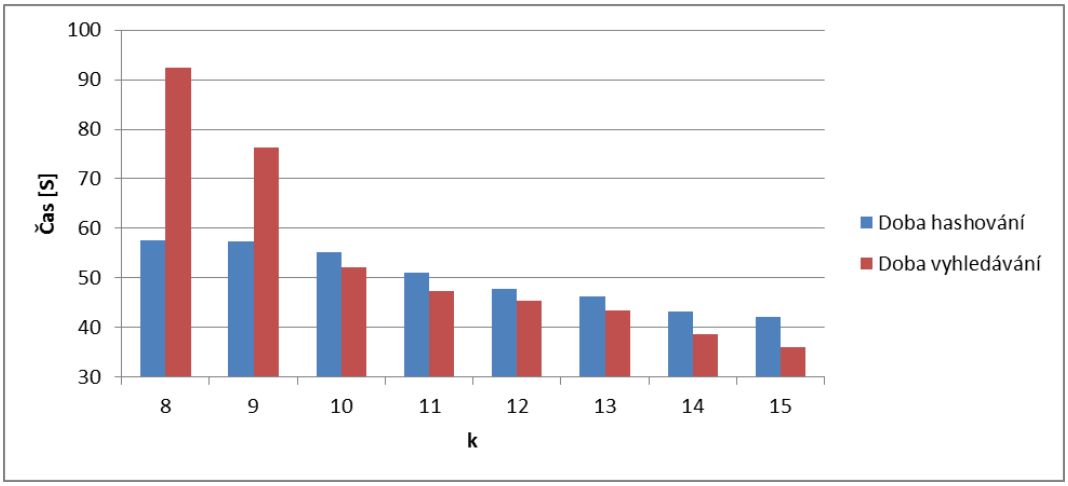
Vstup funkce *tolerance* zajišťuje toleranci pro inzerce a delece. Jde o počet nukleotidů, které mohou být vloženy mezi 2 posloupnosti *k*-merů. V praxi to znamená, že se *shifty* pro dané *i* mohou lišit až o hodnotu tolerance a i přesto jsou zaregistrovány jako posloupnost *k*-merů. Skóre těchto blízkých posloupností je následně sečteno. Následně je vybráno maximální skóre pro dané *i* (tzn. co nejdelší možná posloupnost *k*-merů společná mezi hledanou sekvencí a *i*-tou sekvencí v databázi s tolerancí pro inzerce a delece) a je uloženo do vektoru skóre, kde se postupně shromáždí skóre pro každé *i*.

Jamile je oskórováno každé *i*, je podle skóre vybráno tolik nejlepších výsledků, kolik si uživatel zvolí vstupem *results* a jsou zapsány do dané matice v seznamu výsledků iniciované na začátku od nevyššího skóre po nejnižší (viz příklad níže pro *results =10*). Pokud se skóre shodují, jsou výsledky seřazeny podle pozice sekvence v databázi. V příkladu uvedeném níže lze vidět příklad výpisu nejlepších deseti výsledků pro jednu hledanou sekvenci s názvem *query_sequence_1*.

```
$`query_sequence_1`
      Db number Name                                Score
 [1,] "559"     "IPD:MHC02022 Mafa-A1*004:01"    "225"
 [2,] "868"     "IPD:MHC01801 Mafa-A6*01:01"     "225"
 [3,] "870"     "IPD:MHC02552 Mafa-A6*01:03"     "225"
 [4,] "2350"    "IPD:MHC01862 Mamu-A6*01:01"     "225"
 [5,] "2351"    "IPD:MHC01901 Mamu-A6*01:02"     "225"
 [6,] "2352"    "IPD:MHC01902 Mamu-A6*01:03"     "225"
 [7,] "2353"    "IPD:MHC02411 Mamu-A6*01:04"     "225"
 [8,] "3322"    "IPD:MHC04139 Mane-A6*01:02"     "225"
 [9,] "698"     "IPD:MHC02594 Mafa-A1*071:03"    "196"
[10,] "869"     "IPD:MHC01707 Mafa-A6*01:02"     "196"
```

Po oskórování první hledané sekvence funkce pokračuje ve vyhledávání dalších sekvencí až do konce zadaného FASTA souboru. Výsledkem je seznam, ve kterém, každá položka představuje jednu hledanou sekvenci. V této položce seznamu se nachází tolik nejlepších výsledků pro danou sekvenci, kolik si uživatel zadá na vstupu funkce.

Obrázek 12 znázorňuje časy hashování databáze a vyhledávání v ní pro různá *k*. Databáze představuje 4326 MHC sekvencí DNA o průměrné délce 705 bazí. Hledáno v ní bylo 23 neznámých unikátních sekvencí o průměrné délce 190 bází.

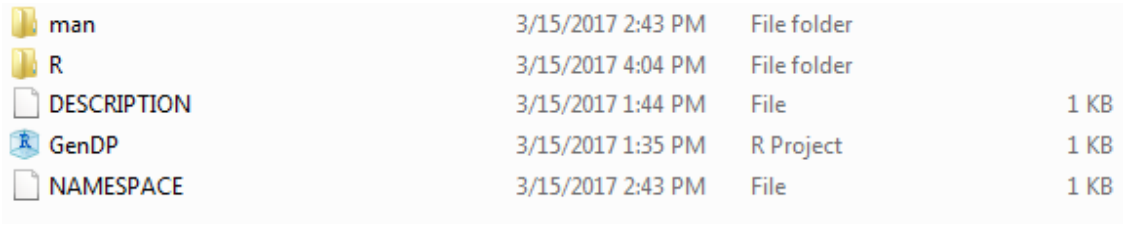


Obrázek 12: Graf časů zahashování databáze a vyhledávání v ní pro různá k

4.6 Generování a distribuce balíčku v R

R v sobě obsahuje propracovaný systém tvorby balíčků funkcí. Jde o jednu z největších výhod tohoto programovacího jazyka. Komunita vývojářů tvoří balíčky, které se navzájem podporují, doplňují a zavádějí nové funkcionality. Pro vystavění balíčku z funkcí je nejjednodušší využít balíček *devtools*, který je dostupný v CRAN.

Balíčky vytvořené v R mají svou pevně danou strukturu. Obrázek 13 znázorňuje, jak by měla vypadat složka s balíčkem. Nejdůležitější je složka *R*, která obsahuje kód funkcí, které mají být součástí balíčku.



Obrázek 13: Adresář R balíčku

Soubor *DESCRIPTION* obsahuje obecné údaje o balíčku. Již v této fázi je možné balíček zabalit do souboru s příponou .tar.gz a distribuovat jej, ale aby byl balíček kompletní, je nezbytné do něj přidat dokumentaci a soubor *NAMESPACE*.

Pro efektivní vývoj a využívání softwaru je přesná a zevrubná dokumentace velice důležitá. Musí být pravidelně udržována a konzistentní. Součástí R je systém pro takovouto dokumentaci, kdy je každá funkce popsána se svými vstupy, výstupy a hlavně příkladem využití. Bioconductor využívá tento systém dokumentace a zároveň do ní zavádí koncept tzv. vinět. Viněty jsou jakousi vyšší úrovní dokumentace, jejímž úkolem je popsat chování a vzájemné interakce skupiny funkcí. Jde o návody, jak s pomocí vícero funkcí dosáhnout určitého výsledku, kterého samostatnými funkcemi dosáhnout nelze.

Dokumentaci lze napsat ručně, ale tento proces je poměrně složitý. Dokumentace pro jednotlivé funkce se zapisuje do textových souborů s příponou .Rd do podsložky balíčku *man* ve formátu podobném LatTexu. Aby nebylo nutné tento formát využívat, dá se v balíčku *devtools* pro dokumentaci využít formát Roxygen2. Roxygen2 je formát dokumentace, který se píše ve formě R komentářů, kterým předchází znak. Roxygen2 využívá různých parametrů k popisu vstupů, výstupů a příkladů využití funkce.

Tento formát má i tu výhodu, že dokumentace se nachází přímo v kódu funkce, takže není nutno ji ukládat odděleně od kódu, jako v nativním formátu R. Příkazem `document()` se poté vygeneruje nativní dokumentace pro jednotlivé funkce R balíčku do složky *man*. Tato dokumentace může být vyexportována ve formátu pdf nebo zobrazena přímo v R. Tento příkaz také vygeneruje soubor *NAMESPACE*, který říká, které z funkcí budou přístupné uživateli balíčku.

V této fázi je již balíček kompletní a je připraven pro distribuci. Lze jej zabalit do formátu tar.gz příkazem `build()`. Další možností je distribuce přes GitHub. GitHub nabízí verzovací systém Git a bezplatný webhosting pro open source projekty. Je nutno vytvořit nový projekt s názvem balíčku (v případě této práce byl zvolen název GenDP) a nahrát do něj celý adresář s vytvořeným balíčkem. Následně může každý uživatel balíček nainstalovat jednoduchým příkazem z balíčku *devtools* `install_github("jmenoUctu/nazevBalicku")`. V případě této práce jde o příkaz `install_github("janmatula/GenDP")`.

5 Vyhodnocení funkčnosti navržených nástrojů

5.1 Funkce demultiplex

Otestování funkce `demultiplex` proběhlo na souborech dat získaných při MHC a KIR sekvenování.

Soubor MHC sekvenačních dat obsahoval 147122 sekvencí rozdělených do devíti vzorků. Počty demultiplexovaných sekvencí pro jednotlivé vzorky v dopředném i reverzním směru znázorňuje Tabulka 7.

Tabulka 7: Demultiplexovaná data z MHC sekvenování

Kód vzorku	Původ vzorku	Počet forward readů	Počet reverse readů	Suma
S1	RSb8 (Slezina)	6891	7027	13918
S2	RCv7 (Slezina)	8980	9349	18329
S3	RKr11 (IngLN)	10159	10329	20488
S5	RKr11 (IngLN)	4805	4500	9305
S6	RKF12 (Slezina)	3371	3134	6505
S7	RKF12 (Slezina)	4457	4139	8596
S9	RJi8 (IngLN)	7788	7994	15782
S10	RMi10 (Slezina)	8281	7367	15648
S11	RFb9 (PBMC)	4835	5652	10487
Suma		59567	59491	119058

Z tabulky lze vyčíst, že celkový počet sekvencí je po demultiplexaci o 28064 nižší než celkový počet sekvencí před demultiplexací, což představuje téměř 20% ztrátu dat. To je způsobeno faktem, že některé multiplexační identifikátory nejsou při sekvenování přečteny správně a takto označené sekvence jsou tedy zanedbány.

Dalším souborem dat, na kterém byla funkce otestována, představují surová data získaná KIR sekvenováním. Celkový počet sekvencí v tomto souboru dat je 111762 Počty sekvencí pro jednotlivé demultiplexované vzorky znázorňuje Tabulka 8.

Tabulka 8: Demultiplexovaná data z KIR sekvenování

Kód vzorku	Původ vzorku	Počet forward readů	Počet reverse readů	Suma
17DH	RSb8 (PBMC)	2819	1916	4735
17DL	RSb8 (PBMC)	3482	2521	6003
20DH	RSb8 (Slezina)	3695	2414	6109
20DL	RSb8 (Slezina)	2939	2038	4977
21DH	RSb8 (IngLN)	3996	2215	6211
21DL	RSb8 (IngLN)	2967	1697	4664
23DH	RJi8 (Slezina)	2584	1467	4051
23DL	RJi8 (Slezina)	3286	1863	5149
25DH	RJi8 (IngLN)	3220	2298	5518
25DL	RJi8 (IngLN)	2255	1703	3958
26DH	RDu8 (AxLN)	3643	2517	6160
26DL	RDu8 (AxLN)	3082	1935	5017
27DH	RDu8 (Slezina)	3111	2154	5265
27DL	RDu8 (Slezina)	2922	2059	4981
29DH	RCv7 (PBMC)	3186	1847	5033
29DL	RCv7 (PBMC)	2875	1876	4751
54DH	RA111 (PBMC)	3196	2169	5365
54DL	RA111 (PBMC)	2516	1907	4423
Suma		55774	36596	92370

Stejně jako u sekvenovaných MHC dat lze z tabulky vidět, že byly při demultiplexaci zanedbány sekvence s neznámým či poškozeným multiplexačním identifikátorem. Konkrétně jde o 19392 sekvencí DNA, což představuje téměř 18 % celkového počtu sekvencí v původním souboru.

5.2 Funkce condense

Pro otestování funkce `condense` jsou využity data, které vznikly demultiplexací KIR a MHC surových sekvenačních dat. Kondenzaci MHC sekvencí znázorňuje Tabulka 9, kondenzaci KIR sekvencí Tabulka 10. Hodnota prahu byla ponechaná v obou případech na hodnotě 10.

Tabulka 9: Kondenzace MHC sekvencí

Kód vzorku	Původ vzorku	Počet forward readů	Počet reverse readů	Suma
S1	RSb8 (Slezina)	32	33	65
S2	RCv7 (Slezina)	38	50	88
S3	RKr11 (IngLN)	38	38	76
S5	RKr11 (IngLN)	26	32	58
S6	RKF12 (Slezina)	23	25	48
S7	RKF12 (Slezina)	20	19	39
S9	RJi8 (IngLN)	36	43	79
S10	RMi10 (Slezina)	36	36	72
S11	RFb9 (PBMC)	72	79	151
Suma		321	355	676

Tabulka 10: Kondenzace KIR sekvencí

Kód vzorku	Původ vzorku	Počet forward readů	Počet reverse readů	Suma
17DH	RSb8 (PBMC)	7	7	14
17DL	RSb8 (PBMC)	11	11	22
20DH	RSb8 (Slezina)	7	7	14
20DL	RSb8 (Slezina)	10	8	18
21DH	RSb8 (IngLN)	10	6	16
21DL	RSb8 (IngLN)	9	7	16
23DH	RJi8 (Slezina)	8	7	15
23DL	RJi8 (Slezina)	5	3	8
25DH	RJi8 (IngLN)	9	8	17
25DL	RJi8 (IngLN)	7	10	17
26DH	RDu8 (AxLN)	14	11	25
26DL	RDu8 (AxLN)	8	7	15
27DH	RDu8 (Slezina)	13	10	23
27DL	RDu8 (Slezina)	8	6	14
29DH	RCv7 (PBMC)	12	6	18
29DL	RCv7 (PBMC)	13	8	21
54DH	RA111 (PBMC)	10	10	20
54DL	RA111 (PBMC)	3	3	6
Suma		164	135	299

V těchto tabulkách lze vidět, že počty sekvencí byly oproti původnímu souboru značně omezeny. Z celkového počtu sekvencí DNA nám po demultiplexaci a

kondenzaci zbylo u MHC pouze přibližně 0,5 % celkového počtu sekvencí a u KIR dokonce jen 0,25 % celkového počtu sekvencí z původního souboru. Jakákoli práce s těmito daty bude tedy nesrovnatelně rychlejší, než kdybychom používali surová sekvenační data.

5.3 Vyhledávání v referenční databázi s pomocí algoritmu SSAHA

Pro otestování funkčnosti vyhledávání byly použity surové sekvence DNA, které byly následně demultiplexovány do vzorků a kondenzovány navrženými funkcemi `demultiplex` a `condense` s prahem 10. Jako referenční databáze byla využita Immuno polymorphism database popsaná v kapitole 3.3.1, která byla stažena z ftp serveru EMBL-EBI navrženou funkcí `downloadDb`. Celkem databáze obsahuje 4326 sekvencí o celkové délce přibližně 3 miliony nukleotidů.

Délka *k*-meru pro zahashování databáze byla zvolena na hodnotu 7. Databáze byla zahashována a následně uložena na disk, kde k ní bylo dále přistupováno. Zahashovaná databáze má velikost 900 kB, což představuje oproti nezahashovanému FASTA souboru o velikost 3173 kB podstatný rozdíl.

V databází bylo vyhledáváno celkem 676 sekvencí DNA rozdělených do devíti vzorků. Délka *k*-meru pro vyhledávání musí být zvolena stejně, jako jeho délka pro zahashování databáze, tzn. 7. Parametr tolerance pro inserce a delece byl ponechán na hodnotě 1. Bylo vypsáno 10 nejlepších výsledků pro každou hledanou sekvenci.

5.3.1 Porovnání navržené metody s metodou BLAST

Pro statistickou analýzu předzpracovaných dat bylo v práci využito srovnání s metodou BLAST. BLAST (Basic Local Alignment Search Tool) je jedním z nejpoužívanějších algoritmů v bioinformatice pro porovnávání biologických sekvencí s databází. Porovnávat se mohou jak proteinové sekvence (BLASTp), tak nukleotidové sekvence (BLASTn). BLAST nachází podobné sekvence tak, že vyhledává krátké shody mezi sekvencemi. Po nálezu první shody BLAST začne vytvářet lokální zarovnání sekvencí, které je ohodnoceno skórem [29].

BLAST byl využíván z příkazového řádku počítače. Nejprve bylo nutno připravit požívanou databázi příkazem `makeblastdb`. Následně bylo provedeno vyhodnocení pro všechny hledané sekvence z MHC sekvenování s tím, že pro každou sekvenci bylo vypsáno 1000 nejlepších výsledků shod s databází. Byla využita matice blastn matrix 1, penalizace mezery -2, match 0 a prodloužení 2,5.

Protože je skórovací systém BLAST jiný než skórování navrženého algoritmu, nebylo možno obě metody porovnávat přímo. Dalším problémem bylo, že mnohdy mělo velké množství sekvencí stejné skóre a nebylo možno určit, která z nich má větší váhu.

Proto byl pro srovnání navržené metody s BLAST navržen parametr *DSC* (Database Search Comparison).

Aby bylo vůbec možné výsledky BLAST a navržené metody srovnat, bylo nejprve nutné načíst výsledky BLAST do R. K tomu byla navržena funkce `readBlastResults`, která načte textový soubor, který je výstupem BLAST (formát výstupu musí být nastaven na –outfmt 0), do R a převede jej do stejného formátu, jaký má výstup navržené metody. Uživatel si může zvolit, kolik výsledků pro každou hledanou sekvenci chce načíst.

Pro výpočet parametru *DSC* byla navržena funkce `compareToBlast`, jejímiž vstupy jsou seznam výsledků algoritmu SSAHA a seznam výsledků algoritmu BLAST. *DSC* vyjadřuje, do jaké míry je prvních 10 výsledků metody SSAHA podobných nejlepším výsledkům metody BLAST u každé hledané sekvence DNA. Je vyjádřen v procentech, kdy 0 % znamená, že výsledky SSAHA se vůbec neshodují s výsledky BLAST a 100% znamená, že prvních 10 výsledků SSAHA je shodných s nejlepšími desíti výsledky BLAST. Funkce hodnotí shodnost v pořadí jednotlivých výsledků mezi SSAHA a BLAST a pozici výsledku vůči odpovídajícímu výsledku BLAST. Za každý rozdíl mezi výsledky metod je odečten určitý počet bodů. *DSC* je potom vyjádřením výsledného počtu bodů vůči maximálnímu možnému počtu bodů v procentech.

V průměru dosahuje metoda SSAHA ve srovnání s BLAST podle navrženého parametru 91,87% úspěšnosti. S pomocí tohoto parametru lze poté sledovat validitu daného výsledku v SSAHA oproti BLAST.

5.3.2 Výsledky metody SSAHA pro MHC sekvence

Tabulka 11 zobrazuje výsledek vyhledání sekvencí DNA v referenční databázi pro vzorek S7. Výsledky jsou rozděleny na čtení, která byla osekvenována v dopředném směru a čtení osekvenována v reverzním směru. Tabulka pro dané sekvence z kondenzovaného FASTA souboru vzorku znázorňuje nejlepší shodu z referenční databáze, jeho skóre a jeho porovnání s BLAST s pomocí popsaného parametru *DSC*. Tabulky výsledků pro ostatní vzorky budou uvedeny v příloze A práce.

Tabulka 11: Výsledky metody SSAHA pro vzorek S7

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
842	Mafa-A1*001:01:02	732	98,75	723	Mafa-A1*001:01:02	732	98,75
706	Mamu-B*047:01	682	100,00	623	Mamu-B*047:01	682	100,00
589	Mamu-B*024:01	678	59,93	454	Mamu-B*024:01	678	59,93
378	Mamu-B*019:01:02	731	99,35	299	Mamu-B*019:01:02	731	99,35
146	Mafa-A1*001:01:02	679	98,86	241	Mamu-B*082:02	735	97,78
137	Mamu-B*047:01	631	100,00	151	Mafa-A1*001:01:02	679	98,86
135	Mamu-B*024:01	627	62,28	132	Mamu-B*047:01	631	100,00
86	Mamu-B*046:01:01	735	99,02	108	Mamu-B*024:01	627	62,28
85	Mamu-B*019:01:02	678	99,38	92	Mamu-B*046:01:01	735	99,02
67	Mafa-B*057:02	683	100,00	58	Mamu-I*01:24	733	100,00
65	Mamu-B*082:02	735	97,78	57	Mamu-B*019:01:02	678	99,38
28	Mamu-B*046:01:01	682	99,07	50	Mamu-B*082:02	682	97,78
27	Mamu-B*051:04	731	99,35	48	Mafa-B*057:02	683	100,00
18	Mamu-I*01:24	733	100,00	23	Mamu-A2*05:04:01	736	100,00
18	Mamu-B*072:01:02	735	99,46	21	Mamu-B*051:04	731	99,35
13	Mafa-A1*001:01:02	488	40,10	18	Mamu-B*072:01:02	735	99,46
13	Mane-B*024:01	578	68,48	17	Mamu-I*01:24	680	100,00
13	Mafa-B*057:02	632	100,00	13	Mamu-A2*05:04:01	683	100,00
12	Mamu-A2*01:01	535	58,73	13	Mamu-I*01:20:02	733	100,00
11	Mane-B*047:01:02	531	53,94				

V levém sloupci tabulky je zobrazen počet identických sekvencí pro daný vzorek, které byly obsaženy v původním souboru surových sekvenačních dat. Ve druhém sloupci je shoda s danou sekvencí DNA z databáze s největším skórem podle metody SSAHA. V případě, že mělo více vzorků stejné skóre, byl z vizualizačních důvodů vybrán pouze jeden z nich a to ten, který byl v databázi nalezen dříve. Následuje sloupec SSAHA skóre (princip jeho výpočtu je popsán v kapitole 4.5.3). Posledním sloupcem tabulky je vypočtený parametr *DSC* pro srovnání s metodou BLAST.

Lze vidět, že navržený nástroj se svou průměrnou 91,87% úspěšností konkuruje BLASTU a zároveň lze využívat přímo v prostředí R, bez nutnosti instalace dalších nástrojů a znalosti práce s příkazovým řádkem. Protože jsou výsledky ve struktuře *list*, je další práce s nimi velice jednoduchá a lze je potom dále zpracovávat a exportovat požadovaným způsobem. Nevýhodou implementace algoritmu do prostředí R je jeho časová náročnost oproti metodě BLAST, a to i přesto, že byla snaha kód v co největší míře vektorizovat.

Závěr

Bakalářská práce se zabývá genotypizací u makaků ve výzkumu infekce virem HIV. Teoretická část podrobně rozebírá sekvenační metody od jejich zrodu k těm nejpokročilejším metodám. Dále rozebírá důležitost genotypizace a sekvenování ve výzkumu infekce virem HIV, kdy diskutuje roli hlavního histokompatibilního komplexu v boji proti infekci. Zde se také krátce diskutuje důležitost biologických databází a je představena použitá databáze IPD. V poslední kapitole teoretické části práce jsou rozebrány techniky sekvenování, kdy je zvláštní důraz kladen na pochopení amplikonového sekvenování a využití multiplexačních identifikátorů v masivně paralelním sekvenování.

Další část práce se již soustředí na tvorbu balíčku nástrojů, které slouží ke zpracování surových sekvenačních dat. V úvodu se stručně objasňuje role programovacího jazyka R a projektu Bioconductor v bioinformatice. Dále je představen nástroj pro demultiplexaci sekvencí DNA ze surových sekvenačních dat s využitím multiplexačních identifikátorů. Dalším nástrojem v balíčku je funkce pro kondenzaci, neboli odstranění opakujících se čtení, těchto demultiplexovaných dat. Dalším vytvořeným nástrojem je funkce pro stažení referenční databáze z ftp serveru EMBL-EBI, která bude následně využita pro identifikaci neznámých sekvencí.

Pro identifikaci neznámých sekvencí DNA podle referenční databáze bylo přistoupeno k využití algoritmu SSAHA (Sequence Search and Alignment by Hashing Algorithm). Je vysvětlen princip tohoto algoritmu a je provedena jeho implementace do R. Poslední část kapitoly se zabývá vytvořením balíčku v R a jeho distribucí.

Poslední kapitola práce se zabývá otestováním navržených nástrojů na reálných datech získaných sekvenováním na platformě Roche 454. U testování funkčnosti implementace algoritmu SSAHA také proběhlo srovnání s nástrojem BLAST a je představen parametr *DSC* pro vyhodnocení úspěšnosti navrženého nástroje oproti metodě BLAST. V souvislosti s tím je také navržena funkce pro načtení textového souboru s výsledky algoritmu BLAST do R a funkce pro výpočet navrženého srovnávacího parametru *DSC*.

Navržený nástroj byl využit k identifikaci celkem 676 sekvencí MHC podle referenční databáze obsahující 4326 sekvencí DNA o celkové délce 3 milionů bází. U každé takto identifikované sekvence byl proveden výpočet parametru *DSC*. V průměru tato identifikace dosahuje podle navrženého parametru *DSC* více než 90% úspěšnosti ve srovnání s BLAST.

Navržený balíček funkcí je veřejně dostupný z GitHub a je jednoduše naistalovatelný příkazem `install_github("janmatula/GenDP")` z balíčku *devtools*.

Literatura

[1] BENSON, Dennis A., et al. GenBank. *Nucleic acids research*, 2000, 28.1: 15-18.

[2] MAXAM, Allan M.; GILBERT, Walter. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 1977, 74.2: 560-564.

[3] ŠMARDA, Jan, Jiří DOŠKAŘ, Roman PANTŮČEK, Vladislava RŮŽIČKOVÁ a Jana KOPTÍKOVÁ. Metody molekulární biologie. 1. vyd. Brno: Masarykova univerzita, 2005. 194 s. 1. vydání. ISBN 80-210-3841-1.

[4] ANSORGE, Wilhelm J. Next-generation DNA sequencing techniques. *New biotechnology*, 2009, 25.4: 195-203.

[5] What is the 454 method of DNA sequencing? [online]. www.yourgenome.org, 2015 – [cit. 12/2015]. Dostupné na <http://www.yourgenome.org/facts/what-is-the-454-method-of-dna-sequencing>.

[6] AHMADIAN, Afshin; EHN, Maria; HOBER, Sophia. Pyrosequencing: history, biochemistry and future. *Clinica chimica acta*, 2006, 363.1: 83-94.

[7] Illumina sequencing technology [online]. Illumina inc., 2010 – [cit. 12/2015]. Dostupné na < www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf >.

[8] METZKER, Michael L. Sequencing technologies—the next generation. *Nature reviews genetics*, 2010, 11.1: 31-46.

[9] Overview of SOLiD™ Sequencing Chemistry [online]. Life Technologies, 2013 – [cit. 12/2015]. Dostupné na <http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html>.

[10] MERRIMAN, Barry, et al. Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis*, 2012, 33.23: 3397-3417.

[11] MCCARTHY, Alice. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & biology*, 2010, 17.7: 675-676.

[12] SCHADT, Eric E.; TURNER, Steve; KASARSKIS, Andrew. A window into third-generation sequencing. *Human molecular genetics*, 2010, 19.R2: R227-R240.

[13] 454 Sequencing System Guidelines for Amplicon Experimental Design [online]. 454 Life Sciences Corp., 2011 – [cit. 12/2015]. Dostupné na <http://my454.com/downloads/my454/applications-info/454SequencingSystem_GuidelinesforAmpliconExperimentalDesign_July2011.pdf >.

[14] ŽÁK, Petr. Nové možnosti v sekvenování. *Roche s.r.o., Diagnostics Division*, 2009.

[15] STUDY, The International HIV Controllers, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science (New York, NY)*, 2010, 330.6010: 1551.

[16] WISEMAN, Roger W., et al. Major histocompatibility complex genotyping with massively parallel pyrosequencing. *Nature medicine*, 2009, 15.11: 1322-1326.

[17] JANEWAY, Charles A., et al. The major histocompatibility complex and its functions. 2001.

[18] BONTROP, Ronald E.; WATKINS, David I. MHC polymorphism: AIDS susceptibility in non-human primates. *Trends in immunology*, 2005, 26.4: 227-233.

[19] VALENTINE, Laura E.; WATKINS, David I. Relevance of studying T cell responses in SIV-infected rhesus macaques. *Trends in microbiology*, 2008, 16.12: 605-611.

[20] ZELNÍKOVÁ, J. Využití deep-sequencing při studiu závislostí průběhu lentivirové infekce na variabilitě receptorů přirozené imunity, *Farmaceutická fakulta v Hradci Králové*, 2014. 114 s.

[21] HUBER, I., et al. Cytogenetic mapping and orientation of the rhesus macaque MHC. *Cytogenetic and genome research*, 2003, 103.1-2: 144-149.

[22] ANZAI, Tatsuya, et al. Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proceedings of the National Academy of Sciences*, 2003, 100.13: 7708-7713.

[23] HSU, Katharine C., et al. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. *Immunological reviews*, 2002, 190.1: 40-52.

[24] Services [online]. EMBL-EBI 2017 – [cit. 4/2017]. Dostupné na <http://www.ebi.ac.uk/services/all>.

[25] ROBINSON, James, et al. IPD—the immuno polymorphism database. *Nucleic acids research*, 2012, gks1140.

[26] GENTLEMAN, Robert C., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 2004, 5.10: R80.

[27] ANDERSON, Isobel; BRASS, Andy. Searching DNA databases for similarities to DNA sequences: when is a match significant?. *Bioinformatics*, 1998, 14.4: 349-356.

[28] NING, Zemin; COX, Anthony J.; MULLIKIN, James C. SSAHA: a fast search method for large DNA databases. *Genome research*, 2001, 11.10: 1725-1729.

[29] ALTSCHUL, Stephen F., et al. Basic local alignment search tool. *Journal of molecular biology*, 1990, 215.3: 403-410.

Seznam symbolů, veličin a zkratek

HIV	human immunodeficiency virus
dNTP	deoxyribonukleotidtrifosfát
ATP	adenosintrifosfát
MID	multiplex identifier
KIR	killer cells immunoglobuline-like recetor
MHC	majot histocompatibility complex
HLA	human leukocyte antigens
AIDS	acquired immune deficiency syndrome
A	adenin
C	cytosin
T	tymin
G	guanin
PCR	polymerázová řetězová reakce
emPCR	emulzní polymerázová řetězová reakce
SSAHA	Sequence Search and Alignment by Hashing Algorithm
BLAST	Basic Local Alignment Search Tool
DSC	Database Search Comparison
IPD	Immuno Polymorphism Database

Seznam příloh

A. Výsledky testování algoritmu SSAHA na vzorcích dat

Tabulka I: Výsledky analýzy SSAHA pro vzorek MHC S1

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
1416	Mane-B*051:04:01	579	97,14	920	Mane-B*051:04:01	579	97,14
899	Pacy-B*03	535	90,48	807	Pacy-B*03	535	90,48
607	Mamu-A1*004:05	536	97,14	771	Mamu-A1*004:05	536	97,14
484	Mane-B*051:04:01	532	97,14	648	Mafa-B*007:03	583	99,33
364	Mafa-B*005:01	578	86,90	309	Mafa-B*005:01	578	86,90
297	Mafa-A1*004:01	582	99,46	288	Mafa-A1*004:01	582	99,46
266	Pacy-B*03	490	90,48	271	Mane-B*051:04:01	453	96,43
259	Mamu-A1*004:05	491	97,33	246	Pacy-B*03	490	90,48
169	Mafa-B*007:03	583	99,33	243	Mamu-A1*004:05	491	97,33
127	Mafa-B*005:01	531	87,74	238	Mane-B*051:04:01	532	97,14
78	Mafa-A1*004:01	535	99,46	207	Mafa-B*007:03	536	100,00
71	Mamu-B*054:02	577	95,76	191	Mane-B*051:04:01	412	96,43
66	Mafa-B*007:03	536	100,00	132	Mafa-B*005:01	531	87,74
51	Mafa-AG*03:02	533	69,73	108	Mafa-A1*004:01	535	99,46
47	Mafa-A1*071:03	258	97,24	58	Mamu-B*054:02	577	95,76
45	Paan-AG*02	634	86,06	41	Mafa-B*109:02	631	99,23
45	Mamu-A1*004:05	284	97,14	36	Mafa-AG*03:02	533	69,73
35	Mamu-A1*004:05	491	96,43	35	Mamu-I*01:01:02	579	100,00
32	Mafa-AG*03:02	488	69,73	34	Mafa-B*109:02	582	98,46
27	Mamu-I*01:01:02	579	100,00	31	Paan-AG*02	634	86,06
21	Mafa-B*046:01:01	111	98,97	28	Mafa-B*057:04	577	100,00
21	Mafa-B*109:02	631	99,23	23	Mamu-B*082:02	489	79,05
21	Mamu-A1*004:05	257	97,33	23	Mamu-B*054:02	530	96,00
20	Mamu-I*01:01:02	532	100,00	23	Mafa-B*098:06	406	36,25
17	Mamu-A1*028:01	630	85,63	21	Mane-B*120:02	538	95,71
16	Paan-AG*02	585	86,06	19	Mafa-A1*071:03	258	97,24
16	Mamu-B*054:02	530	96,00	19	Mamu-B17*01:13	678	100,00
14	Mafa-B*057:04	577	100,00	14	Mafa-B*057:04	530	100,00
14	Mafa-A1*071:03	288	97,33	12	Mamu-B*082:02	446	61,36
14	Mamu-A1*004:05	448	96,77	12	Mafa-A1*071:03	233	97,58
12	Mamu-B*090:01	275	78,21	11	Mafa-B*046:01:01	111	98,98
11	Mafa-B*098:06	406	36,25	11	Mafa-AG*03:02	488	69,73
				11	Mamu-A1*028:01	630	85,63

Tabulka II: Výsledky analýzy SSAHA pro vzorek MHC S2

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
967	Mafa-B*101:01:01	634	98,06	849	Mafa-A1*085:02	633	85,93
898	Mafa-A1*085:02	633	85,93	842	Mafa-B*101:01:01	634	98,06
842	Mafa-B*022:01	487	90,00	765	Mafa-B*022:01	487	90,00
730	Mafa-B*025:01	407	85,00	701	Mafa-B*048:02:02	532	84,44
705	Mafa-A1*002:02	580	99,38	642	Mafa-B*025:01	407	85,00
363	Pacy-B*03	631	84,80	632	Mafa-A1*002:02	580	99,38
237	Mamu-B*019:04	488	51,33	330	Pacy-B*03	631	84,80
223	Mafa-B*101:01:01	585	98,24	252	Mamu-B*019:04	488	51,33
219	Mafa-B*022:01	444	90,00	215	Mafa-B*101:01:01	585	98,24
205	Mafa-A1*085:02	583	85,93	187	Mafa-B*048:02:02	487	87,41
186	Mafa-B*005:01	627	86,90	184	Mafa-B*022:01	444	90,00
180	Mafa-B*025:01	368	94,00	176	Mafa-B*025:01	368	94,55
179	Mafa-A1*002:02	533	99,57	173	Mafa-A1*085:02	583	85,93
177	Mafa-B*048:02:02	532	84,44	163	Mafa-B*005:01	627	86,90
161	Mafa-A3*13:07	583	99,35	158	Mafa-A3*13:07	583	99,35
148	Mamu-B*044:01:02	683	98,00	137	Mamu-B*044:01:02	683	98,00
77	Mafa-B*048:02:02	487	87,41	134	Mamu-B*145:01	531	40,94
73	Pacy-B*03	582	83,70	127	Chsa-B*18:01	580	76,10
67	Mamu-B*019:04	445	51,33	93	Pacy-B*03	582	83,70
59	Mafa-AG*03:02	533	69,73	63	Mamu-B*063:03	631	92,38
58	Mamu-B*063:03	631	92,38	63	Mafa-A1*002:02	533	99,57
54	Chsa-B*18:01	580	76,10	61	Mamu-B*019:04	445	51,33
45	Mafa-B*005:01	578	86,90	51	Mafa-B*005:01	578	86,90
43	Mamu-B*044:01:02	632	98,00	49	Mafa-AG*03:02	533	69,73
41	Mamu-A1*120:01	487	98,18	40	Mamu-B*044:01:02	632	98,00
39	Mafa-A3*13:07	536	99,35	38	Mafa-A1*002:02	273	96,36
37	Mafa-B*063:01	631	92,97	35	Mafa-A3*13:07	536	99,35
28	Mamu-B*063:03	535	97,71	34	Mamu-B*145:01	486	40,94
27	Paan-AG*02	737	91,67	32	Mamu-A1*120:01	487	98,18
27	Mamu-B*145:01	531	40,94	30	Chsa-B*18:01	533	76,10
22	Mamu-I*01:01:02	579	100,00	25	Paan-AG*02	737	91,67
19	Mamu-B*063:03	582	92,38	22	Mamu-B*018:01	492	77,91
18	Mamu-B*015:04:02	534	75,69	21	Mafa-A1*002:02	329	93,78
16	Mafa-B*064:01	579	80,00	20	Mafa-B*063:01	631	92,97
13	Mafa-B*044:04	538	96,52	20	Mamu-I*01:01:02	579	100,00
13	Mafa-I*01:12:01	492	99,20	16	Mafa-B*109:02	631	83,33
12	Mamu-B*018:01	492	77,91	15	Mamu-B*063:03	582	92,38
12	Mafa-B*005:01	292	86,90	15	Mafa-B*044:04	538	96,52
				15	Mamu-A1*028:01	630	85,63
				14	Mafa-B*109:02	682	99,23
				14	Mafa-B*063:01	582	92,97

				14	Paan-AG*02	634	86,06
				14	Mafa-I*01:12:01	492	99,20
				12	Mamu-B*145:01	267	41,74
				12	Mamu-B*015:04:02	534	75,69
				11	Mamu-B*063:03	535	97,71
				11	Mafa-B*072:03	490	97,93
				11	Mafa-A1*002:02	294	92,50
				11	Mafa-AG*03:02	488	69,73
				11	Paan-AG*02	684	91,67

Tabulka III: Výsledky analýzy SSAHA pro vzorek MHC S3

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
1418	Mafa-A1*008:04	679	99,31	1310	Mafa-A1*008:04	679	99,31
848	Mafa-A2*24:01	678	96,30	877	Mafa-B*007:01:01	683	100,00
765	Mamu-A1*003:12	735	100,00	767	Mafa-A2*24:01	678	96,30
706	Mamu-B*015:04:02	734	100,00	636	Mamu-B*001:01:01	731	64,83
657	Mamu-B*001:01:01	731	64,83	623	Mamu-B*015:04:02	734	100,00
371	Mafa-B*060:01	737	99,56	620	Mamu-A1*003:12	735	100,00
337	Mamu-A3*13:03	685	100,00	328	Mamu-A3*13:03	685	100,00
300	Mafa-A1*008:04	628	99,31	287	Mafa-B*060:01	737	99,56
277	Mafa-B*030:03:02	735	100,00	285	Mafa-B*005:01	679	79,31
250	Mafa-B*005:01	679	79,31	266	Mafa-A1*008:04	628	99,31
230	Mafa-B*007:01:01	683	100,00	259	Mafa-B*030:03:02	735	100,00
200	Mamu-B*044:01:02	736	97,78	189	Mafa-A4*01:01	684	98,86
191	Mafa-A2*24:01	627	96,30	187	Mafa-B*007:01:01	632	100,00
185	Mafa-A4*01:01	684	98,86	173	Mamu-B*044:01:02	736	97,78
173	Mamu-A1*003:12	682	100,00	148	Mafa-A2*24:01	627	96,30
170	Mamu-B*015:04:02	681	100,00	144	Mamu-A1*003:12	682	100,00
121	Mamu-B*001:01:01	678	75,71	123	Mamu-B*015:04:02	681	100,00
109	Mamu-B*072:01:02	735	99,46	116	Mamu-B*001:01:01	678	75,71
95	Mafa-B*057:02	683	100,00	115	Mamu-I*01:24	733	100,00
91	Mamu-A3*13:03	634	100,00	93	Mafa-B*057:02	683	100,00
89	Mafa-B*060:01	684	99,56	91	Mafa-B*060:01	684	99,56
59	Mafa-B*005:01	628	80,00	91	Mamu-B*072:01:02	735	99,46
56	Mafa-B*030:03:02	682	97,93	66	Mafa-B*030:03:02	682	97,93
51	Mafa-A4*01:01	633	98,89	61	Mafa-B*005:01	628	80,00
46	Mafa-B*007:01:01	632	100,00	53	Mamu-A3*13:03	634	100,00
42	Chsa-B*15:04	215	45,67	44	Mafa-A4*01:01	633	98,89
42	Mamu-B*044:01:02	683	97,78	40	Chsa-B*15:04	215	45,67
31	Mamu-B*072:01:02	682	99,49	34	Mafa-B*007:03	683	100,00
28	Mamu-B*015:04:01	582	77,00	31	Mamu-B*044:01:02	683	97,78
24	Mamu-B*188:01	735	96,00	29	Mamu-B*188:01	735	96,00
24	Mafa-B*005:01	316	99,29	25	Mafa-B*057:02	632	100,00
21	Mamu-I*01:24	733	100,00	20	Mamu-I*01:24	680	100,00
14	Chsa-B*15:04	188	35,07	18	Mamu-B*072:01:02	682	99,49
11	Mafa-B*030:03:02	632	99,17	16	Mafa-B*109:02	682	99,23
11	Mafa-A1*003:03	370	82,86	13	Mamu-B*015:04:01	582	77,00
11	Mamu-B*001:01:01	678	66,73	11	Mafa-B*070:01:01	733	98,67
11	Mafa-B*060:01	684	99,47	11	Mafa-B*007:03	583	88,46
11	Mafa-B*057:02	632	100,00	11	Mamu-B*001:01:01	678	66,73

Tabulka IV: Výsledky analýzy SSAHA pro vzorek MHC S5

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
1062	Mamu-A1*004:05	734	100,00	896	Mamu-A1*004:05	734	100,00
648	Mamu-B*048:01	678	100,00	595	Mamu-B*048:01	678	100,00
281	Mamu-B*012:01	678	100,00	246	Mamu-B*012:01	678	100,00
209	Mamu-A1*004:05	681	100,00	181	Mamu-B*022:01	683	98,62
193	Mamu-B*041:01	682	91,03	166	Mamu-A1*004:05	681	100,00
168	Mafa-B*030:01:01	682	99,17	151	Mamu-A4*14:09	737	98,38
156	Mane-B*119:01	734	98,52	144	Mamu-B*041:01	682	91,03
139	Mamu-B*048:01	627	100,00	131	Mamu-B*048:01	627	100,00
128	Mamu-A4*14:09	737	98,38	116	Mafa-B*030:01:01	682	99,17
67	Mamu-B*057:01	679	98,52	116	Mane-B*119:01	734	98,52
67	Mamu-B*012:01	627	90,40	48	Mafa-B*064:01	733	99,33
64	Mafa-B*134:02	631	65,53	44	Mamu-B*057:01	679	98,52
58	Mamu-B*022:01	683	98,62	44	Mamu-B*012:01	627	90,40
34	Mamu-B*041:01	631	91,03	39	Mane-B*119:01	550	92,14
34	Mafa-B*030:03:02	682	99,29	32	Mafa-B*064:01	549	93,55
30	Mane-B*119:01	681	98,52	31	Mamu-B*022:01	632	98,62
29	Mafa-B*064:01	733	99,33	26	Mamu-A1*004:05	630	99,20
25	Mamu-B*046:01:01	735	99,02	26	Mamu-B*041:01	631	91,03
22	Mamu-A1*004:05	630	99,20	24	Mafa-B*134:02	631	65,53
22	Mamu-A4*14:09	683	98,38	23	Mafa-B*134:02	456	89,68
19	Mamu-A1*004:05	681	98,67	23	Mamu-B*046:01:01	735	99,02
16	Mamu-A4*14:09	683	96,57	21	Mafa-B*030:03:02	682	99,29
16	Mane-B*119:01	681	100,00	21	Mamu-A4*14:09	683	98,38
13	Mamu-B*012:01	578	95,65	19	Mamu-A4*14:09	683	96,57
11	Mamu-A1*004:05	533	99,17	17	Mamu-B*053:02	683	99,31
11	Mamu-B*057:01	628	98,71	17	Mane-B*119:01	681	98,52
				16	Mamu-A1*004:05	681	98,67
				13	Mamu-B*053:02	545	99,33
				13	Mafa-B*064:01	680	99,33
				12	Mamu-A1*004:05	533	99,17
				11	Mafa-B*098:02	682	100,00
				11	Mamu-I*01:24	733	100,00

Tabulka V: Výsledky analýzy SSAHA pro vzorek MHC S6

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
446	Mamu-A1*004:05	734	100,00	407	Mamu-B*012:01	678	100,00
351	Mamu-B*012:01	678	100,00	319	Mamu-A1*004:05	734	100,00
311	Mane-B*119:01	734	98,52	270	Mane-B*119:01	734	98,52
310	Mafa-A1*001:01:02	732	98,75	266	Mafa-B*030:01:01	682	99,17
285	Mafa-B*030:01:01	682	99,17	249	Mafa-A1*001:01:02	732	98,75
115	Mamu-B*012:01	627	90,40	127	Mamu-B*082:02	735	97,78
72	Mamu-A1*004:05	681	100,00	88	Mamu-B*057:01	679	98,52
69	Mane-B*119:01	681	98,52	84	Mamu-B*012:01	627	90,40
64	Mamu-B*057:01	679	98,52	69	Mamu-A4*14:09	737	98,38
59	Mafa-B*030:03:02	682	99,29	67	Mamu-A1*004:05	681	100,00
58	Mamu-B*046:01:01	735	99,02	53	Mamu-B*046:01:01	735	99,02
51	Mafa-A1*001:01:02	679	98,86	50	Mafa-A1*001:01:02	679	98,86
41	Mamu-B*082:02	735	97,78	47	Mamu-B*053:02	683	99,31
39	Mamu-A4*14:09	737	98,38	45	Mafa-B*030:03:02	682	99,29
30	Mamu-B*012:02	319	97,60	45	Mane-B*119:01	681	98,52
17	Mamu-I*01:24	733	100,00	31	Mamu-B*082:02	682	97,78
17	Mamu-B*012:01	534	75,20	22	Mamu-I*01:24	733	100,00
17	Mamu-A4*14:09	683	98,38	19	Mamu-B*046:01:01	682	99,07
17	Mamu-B*057:01	628	98,71	14	Mafa-B*030:01:01	534	100,00
15	Mamu-B*053:02	683	99,31	13	Mane-B*119:01	681	100,00
15	Mamu-B*012:01	533	86,96	13	Mamu-A2*05:04:01	736	100,00
14	Mamu-B*012:01	578	95,65	13	Mamu-B*038:01	631	90,00
11	Mane-B*119:01	681	100,00	12	Mamu-B*057:01	628	98,71
				11	Mamu-B*074:01	536	88,89
				11	Mamu-B*012:01	533	86,96

Tabulka VI: Výsledky analýzy SSAHA pro vzorek MHC S7

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
842	Mafa-A1*001:01:02	732	98,75	723	Mafa-A1*001:01:02	732	98,75
706	Mamu-B*047:01	682	100,00	623	Mamu-B*047:01	682	100,00
589	Mamu-B*024:01	678	59,93	454	Mamu-B*024:01	678	59,93
378	Mamu-B*019:01:02	731	99,35	299	Mamu-B*019:01:02	731	99,35
146	Mafa-A1*001:01:02	679	98,86	241	Mamu-B*082:02	735	97,78
137	Mamu-B*047:01	631	100,00	151	Mafa-A1*001:01:02	679	98,86
135	Mamu-B*024:01	627	62,28	132	Mamu-B*047:01	631	100,00
86	Mamu-B*046:01:01	735	99,02	108	Mamu-B*024:01	627	62,28
85	Mamu-B*019:01:02	678	99,38	92	Mamu-B*046:01:01	735	99,02
67	Mafa-B*057:02	683	100,00	58	Mamu-I*01:24	733	100,00
65	Mamu-B*082:02	735	97,78	57	Mamu-B*019:01:02	678	99,38
28	Mamu-B*046:01:01	682	99,07	50	Mamu-B*082:02	682	97,78
27	Mamu-B*051:04	731	99,35	48	Mafa-B*057:02	683	100,00
18	Mamu-I*01:24	733	100,00	23	Mamu-A2*05:04:01	736	100,00
18	Mamu-B*072:01:02	735	99,46	21	Mamu-B*051:04	731	99,35
13	Mafa-A1*001:01:02	488	40,10	18	Mamu-B*072:01:02	735	99,46
13	Mane-B*024:01	578	68,48	17	Mamu-I*01:24	680	100,00
13	Mafa-B*057:02	632	100,00	13	Mamu-A2*05:04:01	683	100,00
12	Mamu-A2*01:01	535	58,73	13	Mamu-I*01:20:02	733	100,00
11	Mane-B*047:01:02	531	53,94				

Tabulka VII: Výsledky analýzy SSAHA pro vzorek MHC S9

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
1514	Mamu-B*029:02	734	98,75	859	Mafa-A1*028:02	736	100,00
1013	Mafa-A1*028:02	736	100,00	841	Mamu-B*029:02	734	98,75
779	Mafa-A1*001:01:02	732	98,75	648	Mafa-B*007:01:01	683	100,00
537	Mamu-B*001:01:01	731	64,83	639	Mafa-A1*001:01:02	732	98,75
331	Mamu-B*029:02	681	98,86	498	Mamu-B*001:01:01	731	64,83
255	Mafa-B*017:01	732	97,27	436	Mamu-B*029:02	550	87,03
253	Mafa-B*030:03:02	735	100,00	236	Mafa-B*017:01	732	97,27
196	Mafa-A1*028:02	682	99,39	217	Mafa-B*030:03:02	735	100,00
152	Mafa-B*007:01:01	683	100,00	192	Mamu-B*029:02	681	98,86
151	Mafa-A1*001:01:02	679	98,86	169	Mafa-A1*028:02	682	99,39
100	Mafa-B*060:01	737	99,56	142	Mafa-A1*001:01:02	679	98,86
98	Mamu-B*001:01:01	678	75,71	140	Mafa-B*007:01:01	632	100,00
89	Mafa-A4*01:01	683	97,78	114	Mamu-B*001:01:01	678	75,71
80	Mamu-A2*05:04:01	684	97,14	97	Mamu-B*029:02	505	87,03
62	Mamu-B*072:01:02	735	99,46	84	Mafa-B*060:01	737	99,56
56	Mafa-B*030:03:02	682	97,93	82	Mamu-I*01:24	733	100,00
44	Mafa-B*017:01	679	97,27	75	Mafa-A1*001:01:02	542	93,75
32	Mamu-I*01:24	733	100,00	73	Mamu-A2*05:04:01	684	97,14
32	Mafa-B*007:01:01	632	100,00	70	Mafa-A4*01:01	683	97,78
26	Mamu-B*029:01:01	681	83,92	58	Mamu-B*072:01:02	735	99,46
25	Mafa-A1*001:01:02	628	73,24	50	Mafa-B*030:03:02	682	97,93
25	Mafa-B*060:01	684	99,56	48	Mafa-B*007:03	683	100,00
23	Mafa-A4*01:01	632	97,22	43	Mafa-B*068:01	393	99,35
21	Mamu-B*188:01	735	96,00	39	Mafa-B*017:01	679	97,27
21	Mamu-A2*05:04:01	633	97,14	31	Mamu-B*188:01	735	96,00
20	Mafa-A1*028:02	632	68,39	30	Mamu-B*072:01:02	681	98,89
20	Mafa-A1*028:02	682	100,00	30	Mamu-I*01:02:01	683	100,00
16	Mafa-B*068:01	393	99,35	24	Mafa-B*060:01	684	99,56
15	Mamu-B*001:01:01	678	66,73	23	Mamu-A2*05:04:01	736	100,00
14	Mamu-B*063:03	735	94,29	20	Mamu-B*029:01:01	681	83,92
13	Mamu-I*01:02:01	683	100,00	19	Mafa-A1*001:01:02	497	94,12
12	Mafa-B*007:03	631	100,00	19	Mamu-B*001:01:01	678	66,73
11	Mamu-B*001:01:01	627	64,42	17	Mamu-B*029:02	505	79,35
11	Mamu-B*001:03	580	64,42	17	Mamu-A2*05:04:01	633	97,14
11	Mamu-B*072:01:02	681	98,89	16	Mafa-A4*01:01	632	97,22
11	Mamu-B*072:01:02	682	99,49	16	Mafa-A1*001:01:02	628	73,24
				16	Mafa-A1*028:02	632	68,39
				15	Mafa-A1*028:02	682	100,00
				15	Mafa-B*007:03	631	100,00
				14	Mamu-B*001:01:01	627	64,42
				12	Mamu-B*063:03	735	94,29

Tabulka VIII: Výsledky analýzy SSAHA pro vzorek MHC S10

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
878	Mamu-A1*007:01	682	99,29	742	Mamu-A1*007:01	682	99,29
781	Mafa-A1*008:04	679	99,31	687	Mafa-A1*008:04	679	99,31
709	Mamu-B*002:01	679	99,26	592	Mamu-B*002:01	679	99,26
597	Mafa-A1*001:01:02	732	98,75	514	Mafa-A1*001:01:02	732	98,75
572	Mamu-B*055:01	678	80,78	509	Mamu-B*055:01	678	80,78
483	Mamu-B*052:01	678	91,88	411	Mamu-B*052:01	678	91,88
372	Mamu-B*058:02	683	98,71	343	Mamu-B*058:02	683	98,71
267	Mamu-A3*13:03	685	100,00	267	Mamu-A1*007:01	631	99,29
258	Mamu-A1*007:01	631	99,29	243	Mamu-B*002:01	628	98,52
252	Mafa-A1*008:04	628	99,31	238	Mafa-A1*008:04	628	99,31
243	Mamu-B*002:01	628	98,52	201	Mamu-A3*13:03	685	100,00
190	Mamu-B*055:01	627	81,87	159	Mamu-B*055:01	627	81,87
190	Mafa-A1*001:01:02	679	98,86	156	Mafa-A1*001:01:02	679	98,86
165	Mamu-B*052:01	627	79,90	114	Mamu-B*052:01	627	79,90
98	Mamu-B*058:02	632	98,71	102	Mamu-B*058:02	632	98,71
97	Mamu-A3*13:03	634	100,00	87	Mamu-A2*05:02:01	681	93,10
78	Mamu-A2*05:02:01	681	93,10	63	Mamu-A3*13:03	634	100,00
45	Mamu-A1*119:01	683	92,35	44	Mamu-A1*119:01	683	92,35
33	Mafa-G*02:03	677	100,00	43	Mafa-B*098:02	682	100,00
26	Mamu-A2*05:02:01	630	93,10	30	Mafa-G*02:03	677	100,00
24	Paan-AG*02	634	86,06	27	Mane-B*061:01	328	86,67
23	Mamu-B*079:03	683	99,53	26	Mafa-B*098:02	632	100,00
22	Mane-B*111:01	332	22,85	26	Mamu-A2*05:02:01	630	93,10
21	Mamu-A1*028:01	681	98,24	22	Pacy-B*04	681	98,50
19	Pacy-B*04	681	98,50	17	Mafa-B*060:01	737	99,56
19	Mafa-B*060:01	737	99,56	17	Mamu-B17*01:01:07	626	100,00
19	Mane-B*061:01	328	86,67	16	Mamu-B*079:03	683	99,53
18	Mafa-AG*04:04:01	682	97,58	15	Paan-AG*02	684	91,18
13	Mafa-B*046:07N	680	99,52	15	Paan-AG*02	634	86,06
13	Paan-AG*02	684	91,18	15	Mamu-B*058:02	632	98,06
13	Mane-B*061:01	293	95,14	14	Mamu-B*134:02	449	85,93
12	Mamu-A1*028:01	630	85,63	14	Mafa-B*098:02	631	100,00
12	Mafa-AG*07:01:02	678	85,82	14	Mafa-AG*04:04:01	682	97,58
12	Mafa-B*098:02	682	100,00	12	Mane-B*111:01	332	22,85
12	Mafa-B*050:02	329	61,67	12	Mamu-A2*05:04:01	736	100,00
12	Mafa-B*060:01	684	99,56	11	Mamu-B17*01:13	627	100,00

Tabulka IX: Výsledky analýzy SSAHA pro vzorek MHC S11

Dopředné čtení				Reverzní čtení			
Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]	Počet sekvencí DNA	Název alely přiřazené z databáze	SSAHA skóre	DSC [%]
310	Mamu-A1*007:02	680	92,86	251	Mamu-A1*007:02	680	92,86
180	Mamu-A1*003:12	684	82,86	164	Mamu-A1*003:12	684	82,86
155	Mane-B*119:01	734	98,52	158	Mafa-B*007:01:01	683	100,00
154	Mamu-B*001:01:01	731	64,83	153	Mamu-B*022:01	683	98,62
143	Mafa-AG*05:01:01	735	92,43	134	Mamu-B*001:01:01	731	64,83
132	Mamu-B*012:01	678	100,00	131	Mamu-B*012:01	678	100,00
114	Mafa-B*030:01:01	682	99,17	120	Mane-B*119:01	734	98,52
107	Mamu-A1*007:02	629	92,86	103	Mamu-A1*007:02	629	92,86
99	Mamu-A4*14:09	737	98,38	101	Mafa-AG*05:01:01	735	92,43
97	Mafa-B*030:03:02	735	100,00	98	Mafa-B*098:02	682	100,00
93	Mamu-B*072:01:02	735	99,46	98	Mamu-B*035:01	680	97,69
87	Mafa-B*050:02	191	42,56	98	Mafa-B*030:03:02	735	100,00
84	Mamu-B*188:01	735	96,00	94	Mamu-B*072:01:02	735	99,46
81	Mafa-A4*01:01	736	99,41	90	Mafa-A4*01:01	736	99,41
79	Mafa-G*02:03	677	100,00	88	Mafa-B*030:01:01	682	99,17
78	Caja-G*07:01:01	156	12,27	87	Mafa-B*109:02	630	99,23
72	Mamu-B*012:01	627	90,40	86	Mamu-A4*14:09	737	98,38
66	Mafa-B*057:02	683	100,00	82	Mafa-B*109:06	627	99,23
66	Mane-B*119:01	681	98,52	81	Mamu-B*070:01	683	98,79
65	Mafa-B16*01:01	678	100,00	80	Mafa-B*098:06	536	74,94
63	Mamu-B*057:01	679	98,52	77	Mamu-B*188:01	735	96,00
63	Mafa-B*050:02	294	60,07	74	Mamu-B17*01:01:07	677	100,00
62	Mamu-B*063:03	735	94,29	74	Mamu-I*01:20:02	733	100,00
61	Mamu-A1*003:12	633	82,86	73	Mafa-AG*03:02	631	91,43
60	Mafa-AG*03:02	735	87,76	72	Mafa-G*02:03	677	100,00
59	Mafa-AG*04:01	678	98,33	66	Mamu-B*057:01	679	98,52
59	Mafa-AG*05:01:01	682	96,84	66	Mafa-AG*05:01:01	682	96,84
58	Mafa-AG*03:02	631	91,43	63	Mamu-B17*01:13	678	100,00
57	Mafa-B16*01:07	678	100,00	62	Mamu-A1*003:12	633	82,86
50	Mafa-B*030:03:02	682	97,93	62	Mamu-I*01:24	733	100,00
47	Mamu-A4*14:09	683	98,38	62	Mafa-B*007:01:01	632	100,00
47	Mamu-B*001:01:01	678	75,71	62	Mamu-B*022:01	632	98,62
46	Mafa-B*030:03:02	682	99,29	58	Mamu-B*053:02	683	99,31
44	Mafa-B*007:01:01	683	100,00	58	Mafa-AG*04:01	678	98,33
43	Mamu-B*072:01:02	682	99,49	55	Mamu-B*012:01	627	90,40
42	Mamu-B*022:01	683	98,62	54	Mafa-B16*01:07	678	100,00
39	Mafa-B*057:02	632	100,00	54	Mamu-B*001:01:01	678	75,71
38	Mafa-B16*01:07	627	100,00	52	Caja-G*07:01:01	156	12,27
38	Mafa-A4*01:01	683	99,39	52	Mamu-B*063:03	735	94,29
37	Mamu-B*063:03	631	98,42	49	Mane-B*119:01	681	98,52
35	Mamu-B*188:01	682	96,07	49	Mafa-B*057:02	683	100,00

28	Mafa-G*02:03	626	100,00	48	Mafa-B*050:02	191	42,56
26	Mafa-B*109:02	630	99,23	46	Mafa-B*050:02	294	60,07
26	Mamu-B17*01:01:07	677	100,00	44	Mafa-AG*03:02	735	87,76
25	Mafa-B*109:06	627	99,23	43	Mamu-B*035:01	629	97,69
25	Mamu-B*035:01	542	97,78	43	Mafa-B*030:03:02	682	99,29
24	Mamu-I*01:20:02	733	100,00	42	Mafa-B16*01:01	678	100,00
24	Mafa-B*050:02	166	76,53	41	Mafa-B*098:06	491	74,94
23	Mamu-B*063:03	682	94,29	37	Mafa-B*109:02	581	98,57
23	Mafa-AG*04:01	627	98,33	36	Mafa-B*098:02	631	100,00
22	Mamu-G*02:02	441	87,03	35	Mamu-B*063:03	631	98,42
22	Mamu-B*057:01	628	98,71	34	Mamu-B*072:01:02	682	99,49
21	Mafa-AG*03:02	682	88,09	34	Mamu-A4*14:09	683	98,38
20	Mamu-A1*065:02	104	62,98	34	Mamu-B17*01:01:07	626	100,00
20	Mamu-B*031:01	682	79,20	33	Mafa-B*109:06	578	98,46
20	Caja-G*07:01:01	143	49,17	29	Mamu-B*070:01	632	98,79
20	Mamu-B*022:01	632	98,62	29	Mafa-B*030:03:02	682	97,93
19	Mamu-B*053:02	683	99,31	28	Mafa-A4*01:01	683	99,39
18	Mafa-B*050:02	261	70,63	28	Mafa-G*02:03	626	100,00
18	Mamu-B*035:01	629	97,69	27	Mafa-B*050:02	166	76,94
16	Mamu-I*01:24	733	100,00	26	Mafa-AG*03:02	582	91,43
14	Mafa-B16*01:01	627	100,00	26	Mamu-I*01:24	680	100,00
14	Mafa-AG*03:02	582	91,43	26	Mamu-G*02:02	441	87,03
13	Mafa-B*007:01:01	632	100,00	25	Mamu-B*188:01	682	96,07
12	Mafa-B*093:01	136	21,40	24	Mafa-B*057:02	632	100,00
12	Mamu-B*070:01	683	98,79	22	Mafa-B16*01:01	627	100,00
12	Mamu-B*063:03	582	98,42	21	Mamu-B*053:02	632	99,31
11	Mamu-B*012:02	319	97,60	20	Mamu-B17*01:13	627	100,00
11	Mafa-B*098:06	384	76,25	20	Mamu-I*01:20:02	680	100,00
11	Mafa-B*078:04	91	67,01	19	Mamu-A1*065:02	104	62,98
11	Mamu-I*01:24	680	100,00	19	Mafa-AG*03:02	682	88,09
11	Mamu-B*031:01	630	79,20	19	Mamu-B*031:01	682	79,20
				18	Mamu-B*057:01	628	98,71
				17	Caja-G*07:01:01	143	49,30
				16	Mamu-B*063:03	682	94,29
				15	Mafa-B16*01:07	627	100,00
				14	Mamu-B*063:03	582	98,42
				13	Mafa-B*050:02	261	71,51
				12	Mafa-AG*04:01	627	98,33

B. Obsah přiloženého CD

- elektronická verze práce ve formátu *pdf*
- **package/** - složka obsahující vytvořený balíček GenDP, soubor *readme.txt* s instrukcemi pro instalaci a uživatelský manuál ve formátu *pdf*
- **data/** - složka obsahující předzpracované MHC a KIR sekvence s rozpisy jednotlivých vzorků a použitou referenční databázi pro MHC